Berechnung singulärer Punkte nichtlinearer Gleichungssysteme

D I S S E R T A T I O N

zur Erlangung des akademischen Grades

Doctor rerum naturalium
(Dr. rer. nat.)

vorgelegt

der Fakultät Mathematik und Naturwissenschaften
der Technischen Universität Dresden

von

Dipl.-Math. Schnabel, Uwe

geboren am 25.09.1968 in Dresden

Gutachter: Prof. Dr. rer. nat. habil. Hubert Schwetlick
Prof. Ph. D. Andreas Griewank
Prof. Ph. D. Eugene Allgower

Eingereicht am: 10.01.2000

Tag der Verteidigung: 27.10.2000

Mein Dank gilt zuerst Herrn Dr. G. Pönisch für seine vielfältigen Hilfen, Hinweise, Anregungen und Korrekturen während des gesamten Promotionsstudiums. Bei der Dissertation gilt er speziell für die Formulierungs- und Gestaltungshilfen zum Kapitel 2. Daneben möchte ich mich auch bei Herrn Prof. Dr. H. Schwetlick für seine Hinweise während der gesamten Zeit bedanken. Außerdem ermöglichte er mir eine Tätigkeit als wissenschaftliche Hilfskraft im Rahmen der DFG-Forschergruppe „Identifikation und Optimierung komplexer Modelle auf der Basis analytischer Sensitivitätsberechnungen“. Für ihre Hilfen bei ADOL-C bedanke ich mich bei Herrn Prof. A. Griewank, Herrn J. Utke und Frau Dr. A. Walther und für Hilfen bei sonstigen Computerproblemen insbesondere bei Herrn R. Lösche und Herrn Dr. T. Schütze. Frau Dr. A. Walther gab außerdem eine Reihe von Formulierungshilfen für das Kapitel 4. Bei Frau Dr. M. Bastian möchte ich mich für ihre Hinweise zum Promotionsverfahren bedanken.

Vielen weiteren ehemaligen und jetzigen Angehörigen der Fachrichtung Mathematik und speziell des Instituts für Numerische Mathematik möchte ich meinen Dank für das gute Arbeitsklima und für kleinere Hilfen im mathematischen Alltag aussprechen.

Inhaltsverzeichnis

Bezeichnungen

In dieser Übersicht stehen zuerst die lateinischen Großbuchstaben, anschließend die lateinischen Kleinbuchstaben, dann die griechischen Buchstaben und zum Schluß die sonstigen Bezeichnungen. Die Buchstaben sind jeweils alphabetisch geordnet. Buchstaben mit Indizes stehen bei den entsprechenden Buchstaben. Darüberhinaus verwendete Indizes werden bei den sonstigen Bezeichnungen aufgeführt. Bezeichnungen, die nur lokal verwendet und dort erklärt werden, werden hier im Regelfall nicht mit aufgenommen. Dies gilt auch für lokale Bezeichnungen, die an anderer Stelle in einem anderen Sinne vorkommen.

A	$A := \partial_{(x,\vartheta)} F(x,\lambda,\alpha) \in \mathbb{R}^{m\times(n+q-l)}$ im Kapitel 3
B, B_i	geränderte Matrix(funktion), vergleiche (2.65), (2.66), (2.67), (3.18), (3.32) und (3.33)
$\boldsymbol{B}$	andere geränderte Matrix(funktion), siehe (4.7)
C, $C_i^{i_1,i_2,j}$	allgemeine Matrizen, $C_i^{i_1,i_2,j} \in \mathbb{R}^{i_1\times i_2}$ kommt im Unterabschnitt 2.2.6 beim Beweis von $g\,(R3)\,\vartheta$ vor
D, $\boldsymbol{D}$, $\bar{D}$	rechte Ränderungen, $D, \boldsymbol{D} \in \mathbb{R}^{m\times q}$, $\bar{D} \in \mathbb{R}^{m\times l}$, siehe (2.26) und (2.9)
$\hat{D}$	$\hat{D} \in \mathbb{R}^{(n+q)\times l}$ siehe (3.21)
$\tilde{D}$	im Kapitel 3 siehe (3.22), sonst siehe (2.34), Verwendung gemäß (2.33)
$\breve{D}$	$\breve{D}^{(i+1)} \in \mathbb{R}^{m\times l}$ siehe Algorithmus 3.46
D_i, D_∞, $D_{2,2}$	Operatoren, siehe Definition 2.30, insbesondere die Definitionen von $g\,(R3)\,h$, $g\,(R4)\,h$ und $g\,(R5)\,h$
E_i	Vertauschungsmatrix, siehe Definition 3.43
F	$F : \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^k \to \mathbb{R}^m : (x,\lambda,\alpha) \mapsto F(x,\lambda,\alpha)$, nichtlineare Funktion
$\boldsymbol{F}$	$\boldsymbol{F}(x,\vartheta,\tau) := F(x,\lambda,\alpha) + \bar{D}\,\mu$, siehe (2.48)
G	$G : \mathbb{R}^{m_{ges}} \to \mathbb{R}^{m_{ges}} : y \mapsto G(y)$, Funktion im erweiterten System, siehe Definition 2.3
H	geeigneter linearer Raum, siehe (3.45)
I_j, I	Einheitsmatrizen $I_j \in \mathbb{R}^{j\times j}$, I geeigneter Dimension
$\tilde{J}$, $\tilde{K}$, J, K	$\tilde{J}, \tilde{K}^T, J, K^T \in \mathbb{R}^{l\times(n-m+q)}$ siehe (3.22) bzw. die Algorithmen 3.27, 3.44 und 3.46
$\breve{J}$, $\breve{K}$	$\breve{J}^{(i+1)}, (\breve{K}^{(i+1)})^T \in \mathbb{R}^{l\times(n-m+q)}$ siehe Algorithmus 3.46
$(K^{(i+1)})_{\iota,j}$, $(J^{(i+1)})_{\iota,j}$	im Unterabschnitt 3.2.7 das Element aus der ι-ten Zeile und der j-ten Spalte von $K^{(i+1)}$ bzw. von $J^{(i+1)}$
$M_{i,j}$	$\begin{bmatrix} I_j \\ 0 \end{bmatrix} \in \mathbb{R}^{i\times j},\ i \geq j$

$N_{i,j}$	$\begin{bmatrix} 0 \\ I_j \end{bmatrix} \in \mathbb{R}^{i\times j},\ i \geq j$
$P \in \mathbb{R}^{n\times n}$	Projektor, einschließlich seiner Matrixdarstellung, siehe (2.6)
P_r	Spalten bilden eine Basis des Bildraums von P, siehe (2.11)
P_s	Spalten bilden eine Basis des Nullraums von P, siehe (2.11)
$Q \in \mathbb{R}^{m\times m}$	Projektor, einschließlich seiner Matrixdarstellung, in Bemerkung 4.3 unbekannte Konstante
Q_r, Q_s	wie P_r und P_s, aber für den Projektor Q, siehe (2.11)
R_u, R_v	$R_u \in \mathbb{R}^{q\times q}$, $R_v \in \mathbb{R}^{(n-m+q)\times(n-m+q)}$, reguläre Matrizen (Koordinatentransformation)
R_1, R_2	$R_1^{(i+1)} \in \mathbb{R}^{(n-m+q)\times(n-m+q)}$, $R_2^{(i+1)} \in \mathbb{R}^{l\times l}$, reguläre Matrizen aus QR-Zerlegungen, siehe Algorithmen 3.27 bzw. 3.46
$\breve{R}_1$, $\breve{R}_2$	$\breve{R}_1^{(i+1)} \in \mathbb{R}^{(n-m+q)\times(n-m+q)}$, $\breve{R}_2^{(i+1)} \in \mathbb{R}^{l\times l}$ reguläre Matrizen, siehe Algorithmus 3.46
$R1, \ldots, R6$	in Definition 2.30 definierte Relationen
S, $\boldsymbol{S}$	siehe Definition 2.22 bzw. in den Abschnitten 4.1 und 4.2 siehe (4.6)
$S_{i,j}$	siehe (2.57)
T, $\boldsymbol{T}$	$T \in \mathbb{R}^{(n-m+q)\times n}$, untere Ränderungen, siehe (2.27), $\boldsymbol{T}$ wird einerseits im Sinne von z. B. (2.38), andererseits im Sinne von (2.54) und (2.55) verwendet
T_0, T_1	$T_0 := M^T_{n+q,n+q-l}\boldsymbol{T}^T \in \mathbb{R}^{(n+q-l)\times(n-m+q)}$ und $T_1 := N^T_{n+q,l}\boldsymbol{T}^T \in \mathbb{R}^{l\times(n-m+q)}$, vergleiche Definition 3.13 mit $\boldsymbol{T} \in \mathbb{R}^{(n-m+q)\times(n+q)}$ aus (2.54)
$\tilde{T}_0$	$\tilde{T}_0 \in \mathbb{R}^{(n+q-l)\times(n-m+q)}$ siehe (3.22)
$\breve{T}_0$	$\breve{T}_0^{(i+1)} \in \mathbb{R}^{(n+q-l)\times(n-m+q)}$ siehe Algorithmus 3.46
$\tilde{\boldsymbol{T}}$	$\tilde{\boldsymbol{T}} \in \mathbb{R}^{(n-m+q)\times(n+q)}$ siehe (3.21) und Bemerkung 3.28 (i)
T_2	$T_2 = \boldsymbol{T}\,N_{n+q,q}$, vergleiche (2.54)
$\bar{T}^{(k)}$	$\bar{T}^{(k)} \in \mathbb{R}^{(n+q)\times(n+q-m)}$ siehe Bemerkung 4.3
U, $\bar{U}$, V, $\bar{V}$	$U, \bar{U} \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{(n+q-l)\times(n+q-l)}$, $\bar{V} \in \mathbb{R}^{(n+q)\times(n+q)}$ orthogonale Matrizen aus der Singulärwertzerlegung von A bzw. $\begin{bmatrix} A & \bar{D} \end{bmatrix}$, siehe Definition 3.2, in den Abschnitten 4.1 und 4.2 $V \in \mathbb{R}^{(n+q)\times(n-m+q)}$ siehe (4.4)
U_0, U_4	zu Σ_0 bzw. Σ_4 gehörenden Spalten von U, vergleiche Definition 3.13
U_r, U_s	Spalten bilden orthonormale Basen des Bildraums von $\partial_x F(x^*, \lambda^*, \alpha^*)$ und seines orthogonalen Komplements, vergleiche (2.4) und (2.5), im Kapitel 3 von $\partial_{(x,\theta)} F(x, \lambda, \alpha)$
V_0, V_4	zu Σ_0 bzw. Σ_4 gehörenden Spalten von V, vergleiche Definition 3.13
V_s, V_r	Spalten bilden orthonormale Basen des Nullraums von $\partial_x F(x^*, \lambda^*, \alpha^*)$ und seines orthogonalen Komplements, vergleiche (2.4) und (2.5), im Kapitel 3 von $\partial_{(x,\theta)} F(x, \lambda, \alpha)$
V_2	siehe (2.17) und (2.18)
W	in den Abschnitten 4.1 und 4.2 siehe (4.4)

W_2	$W_2 \in \mathbb{R}^{(m-q)\times m}$ mit $\ker W_2 = \operatorname{im} Q_s$, wird für Ljapunov-Schmidt-Reduktionen benötigt, siehe (2.17) und (2.18)
X	$X := U^T \bar{D} \in \mathbb{R}^{m\times l}$, vergleiche Definition 3.2
X_i, $\tilde{X}_i$	zu Σ_i gehörende Zeilen von X bzw. von $U^T\tilde{D}$
X_s	$X_s := U_s^T \bar{D}$, vergleiche Definition 3.13
Y	$Y := \bar{V}^T \boldsymbol{T}^T \in \mathbb{R}^{(n+q)\times(n-m+q)}$, vergleiche Definition 3.2
Y_i	zu $\bar{\Sigma}_i$ gehörende Zeilen von Y
Z_i, $\tilde{Z}_i$	zu Σ_i gehörende Zeilen von $V^T T_0$ bzw. von $V^T\tilde{T}_0$
Z_s	$Z_s := V_s^T T_0$, vergleiche Definition 3.13
a, a_i	Richtungsvektoren
c	$c \in \mathbb{R}^m$, siehe (2.35), (2.37), (2.50) (2.53) und (2.55), Parametrisierung
d	$d \in \operatorname{im} \bar{D}$ in (3.19)
e^i	i-te Koordinatenvektor entsprechender Dimension
f	$f : \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^k \to \mathbb{R}^{m_{ges}-m} : (x,\lambda,\alpha) \mapsto f(x,\lambda,\alpha)$, charakterisierende Funktion im erweiterten System
$\boldsymbol{f}$	$\boldsymbol{f} : \mathbb{R}^n \times \mathbb{R}^q \times \mathbb{R}^{p+k+l-q} \to \mathbb{R}^{m_{ges}-m} : (x,\vartheta,\tau) \mapsto \boldsymbol{f}(x,\vartheta,\tau)$, entspricht f, aber mit anderen Variablen, wegen (2.61) auch mit den Argumenten $(\boldsymbol{\xi},\tau,c)$
g, $\boldsymbol{g}$	(verallgemeinerte) Ljapunov-Schmidt-reduzierte Funktionen bzw. deren Bildraumvariable, siehe z. B. (2.33), (2.35), (2.36), (2.49) und (2.50)
h	Hilfsfunktion, siehe (2.46), in einigen Beispielen im Abschnitt 4.3 Diskretisierungsschrittweite
i, j	Indizes
$k \geq 0$	Dimension von α
l	$\operatorname{rank} \partial F(x^*,\lambda^*,\alpha^*) = m - l$ und somit $0 \leq l \leq q$
$\ell \geq 1$	Vielfachheit, z. B. von Rückkehrpunkten, analog als Exponent bzw. Index verwendet
$m > 0$	F bildet in den $\mathbb{R}^m$ ab
m_{ges}	$m_{ges} \geq n+p+k+l \geq m$, Urbild- und Bilddimension im erweiterten System, siehe Definition 2.3
$n > 0$	Dimension von x
$p \geq 0$	Dimension von λ
q	$\operatorname{rank} \partial_x F(x^*,\lambda^*,\alpha^*) = m - q$ mit $\max\{1, m-n+1\} \leq q \leq m$
$r \in \mathbb{R}$	wird im Unterabschnitt 2.2.4 als Dimension von ϑ verwendet und in den Unterabschnitt 3.2.4, 3.2.6 und 3.2.7 $r := \operatorname{rank}(A)$
s, $\boldsymbol{s}$	im Abschnitt 4.2 $s := (x,\vartheta)$ und $\boldsymbol{s} := (\boldsymbol{x},\boldsymbol{\vartheta})$
$t, t_i \in \mathbb{R}$	Koeffizient
u	$u \in \operatorname{im} U_4$ in (3.19)
u^i	i-te Spalte von U, Linkssingulärvektor zum Singulärwert σ_i
v, w	$x = v + w$ z. B. in (2.12), (2.14), (2.15) und (2.16), v kommt in anderer Bedeutung auch in (2.24) und (2.25) vor
v_1, w_1	siehe (2.17) und (2.18)
x, $\boldsymbol{x}$	$x \in \mathbb{R}^n$, Vektor der Zustandsvariablen bzw. Funktionen, die in diesen Raum abbilden

y	Vektor aller Variablen im erweiterten System, $y \in \mathbb{R}^{m_{ges}}$, siehe Definition 2.3
z	Hilfsvektor geeigneter Dimension, z. B. in Lemma 3.6, Beweis von Satz 3.7
z_i	zu Σ_i gehörende Komponenten von z
α	Parameter, die nicht zu λ gehören, $\alpha \in \mathbb{R}^k$
β	siehe (2.24) und (2.25)
γ	$\gamma \in \mathbb{R}^{n+p+k+l-m}$, siehe (4.5) und (4.8)
δ, $\boldsymbol{\delta}$	$\delta, \boldsymbol{\delta} \in \mathbb{R}^{(n-m+q)\times q}$, rechte untere Ränderung
ζ, ζ_i	im Kapitel 3 $\zeta := \measuredangle(\operatorname{im} T_0, \operatorname{im} V_4)$, $\zeta_i := \measuredangle(\operatorname{im} T_0^{(i)}, \operatorname{im} V_4)$, sonst $\zeta = (\zeta_1, \zeta_2) \in \mathbb{R}^2$ zwei Komponenten von (ξ, τ)
$\tilde{\zeta}$	$\tilde{\zeta} := \measuredangle(\operatorname{im} \tilde{T}_0, \operatorname{im} V_4)$
$\eta \in \mathbb{R}^{m-q}$	Parametrisierung des nicht von ξ parametrisierten Teils von x, siehe (2.28)
ϑ, $\boldsymbol{\vartheta}$	$\vartheta \in \mathbb{R}^q$, Parameter bzw. Funktionen, die in diesen Raum abbilden, siehe (2.48), (2.52), (2.53) und (2.55), im Unterabschnitt 2.2.4 wird ein allgemeineres $\vartheta \in \mathbb{R}^r$ betrachtet
$\theta \in \mathbb{R}^{q-l}$	die nicht zu μ gehörenden Komponenten von ϑ
ι	Index
κ	$0 \leq \kappa := \sigma_{m-l+1}/\sigma_{m-l} \leq 1$ bei $0 < l < m$, vergleiche Lemma 3.20
λ, $\boldsymbol{\lambda}$	$\lambda \in \mathbb{R}^p$, ausgezeichneter Parameter bzw. der Vektor der ausgezeichneten Parameter, auch Bifurkations- bzw. Verzweigungsparameter genannt, als Teil von ϑ bzw. $\boldsymbol{\vartheta}$ auch Funktionen
Λ	Funktion, bildet in den Raum von λ ab, siehe Bemerkung 2.23
μ, $\boldsymbol{\mu}$	zusätzlicher Parameter bzw. Vektor der zusätzlichen Parameter im erweiterten System (2.10), $\mu \in \mathbb{R}^l$, als Teil von ϑ bzw. $\boldsymbol{\vartheta}$ auch Funktionen, in anderer Bedeutung in (2.24) und (2.25)
ν	$\nu \in \mathbb{R}^q$, siehe (2.42) und (2.43)
ξ, $\boldsymbol{\xi}$	im Kapitel 3 $\xi := \measuredangle(\operatorname{im} \bar{D}, \operatorname{im} U_4)$, sonst $\xi, \boldsymbol{\xi} \in \mathbb{R}^{n-m+q}$, Parametrisierung eines Anteils von x, siehe (2.28), (2.36) und vergleichbare Systeme
$\tilde{\xi}$, ξ_i	$\tilde{\xi} := \measuredangle(\operatorname{im} \tilde{D}, \operatorname{im} U_4)$, $\xi_i := \measuredangle(\operatorname{im} \bar{D}^{(i)}, \operatorname{im} U_4)$
Ξ, $\boldsymbol{\Xi}$	Funktion, bildet in den Raum von ξ bzw. $\boldsymbol{\xi}$ ab, siehe Definition 2.22
ρ_i, $\tilde{\rho}_i$	siehe (3.39)
ϱ, $\sigma \in \mathbb{R}$	werden in Definition 2.30 zur Definition von D_∞ und $D_{2,2}$ als Variable verwendet, analog in den Unterabschnitten 2.2.4 und 2.2.6 und im Abschnitt 2.3, im Kapitel 3 bezeichnet σ einen Singulärwert, im Abschnitt 4.2 ist $\varrho := (\boldsymbol{\xi}, \tau, c)$
$\sigma_i(C)$	i-ter Singulärwert von C
σ_i, $\bar{\sigma}_i$, $\hat{\sigma}_i$	im Kapitel 3 $\sigma_i := \sigma_i(A)$, $\bar{\sigma}_i := \sigma_i\left(\begin{bmatrix} A & \bar{D} \end{bmatrix}\right)$ und $\hat{\sigma}_i := \sigma_i\left(\begin{bmatrix} A & \bar{D} \\ \boldsymbol{T} & \end{bmatrix}\right)$, vergleiche Definition 3.2

$\check{\sigma}$	für $0 < l < m$ der Wert $\|_r J^{(1)}\|_2$, siehe Algorithmus 3.40, in Bemerkung 3.52 für $l = 0$ verallgemeinert eine Zahl zwischen σ_{m-l} und σ_1, für $l = m$ in Bemerkung 3.42 (iii) $\check{\sigma} \geq \|A\| = \sigma_1$
Σ, $\bar{\Sigma}$	siehe Definition 3.2, enthalten auf der Diagonalen die Singulärwerte von A bzw. $\begin{bmatrix} A & \bar{D} \end{bmatrix}$
Σ_i, $\bar{\Sigma}_i$	Teilmatrizen von Σ bzw. $\bar{\Sigma}$, siehe Definitionen 3.2 und 3.13 und das Lemma 3.23
τ	$\tau \in \mathbb{R}^{p+k+l-q}$, diejenigen Parameter (λ, α, μ), die nicht zu ϑ gehören, siehe (2.48)
Φ	im Abschnitt 4.2 siehe (4.10)
Φ^{j+1}	Taylorkoeffizient von Φ, siehe (4.12)
0	Zahl Null oder Nullmatrix geeigneter Dimension
$0_{\mathbb{R}^i}$	Nullvektor mit $0_{\mathbb{R}^i} \in \mathbb{R}^i$
$^{(i)}$	kennzeichnet die i-te Iterierte, siehe Algorithmen 3.27, 3.40, 3.44 und 3.46 bzw. Abschnitt 4.1
*	kennzeichnet den singulären Punkt
$^+$	kennzeichnet die Moore-Penrose-Pseudoinverse bzw. im Abschnitt 4.1 die neue Iterierte
$^\perp$	kennzeichnet das orthogonale Komplement
$_{beg}$, $_{neu}$	siehe Unterabschnitt 3.2.4, andere Dimensionen als ohne $_{beg}$ bzw. $_{neu}$, auch im Unterabschnitt 3.2.6 verwendet
$_r$, $_s$	siehe Algorithmus 3.40 im Unterabschnitt 3.2.4, kennzeichnet den regulären bzw. den Nullraum-, d. h. den singulären Anteil
$\|\cdot\|$	sofern nicht anders angegeben $\|\cdot\|_2$, d. h. die Euklidische bzw. bei Matrizen die Spektralnorm

Kapitel 1

Einleitung

Viele Probleme lassen sich mathematisch mittels parameterabhängigen nichtlinearen Gleichungssystemen der Form

$$F(x, \lambda, \alpha) = 0 \tag{1.1}$$

mit der nichtlinearen hinreichend glatten Funktion

$$F : \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^k \to \mathbb{R}^m : (x, \lambda, \alpha) \mapsto F(x, \lambda, \alpha)$$

beschreiben. Dies umfaßt auch die Fälle, in denen die Parameter bzw. Parametervektoren λ bzw. α nicht existieren. In diesen Fällen ist $p = 0$ bzw. $k = 0$. Die Aufspaltung der Parameter ergibt sich häufig aus der Problemstellung. Dabei ist λ ein ausgezeichneter Parameter bzw. seltener ein Vektor ausgezeichneter Parameter. Die übrigen Parameter werden dann in α zusammengefaßt. Die Aufspaltung kann beispielsweise danach erfolgen, daß in λ die leicht veränderlichen Parameter zusammengefaßt werden. In diesem Fall gehören zu α dann z. B. solche Parameter, die sich nur durch Änderung des Systemaufbaus bzw. von Systemteilen ändern lassen. In vielen Fällen, z. B. bei der Diskretisierung partieller Differentialgleichungen der Form $\triangle u = f(u, \lambda, \alpha)$, ist dabei $m = n$. Sollte in diesem Fall für einen Lösungspunkt von (1.1) die Jacobimatrix $\partial_x F$ regulär sein, läßt sich dieser Punkt lokal z. B. mit dem Newtonverfahren unter Ausnutzung der überlinearen Konvergenz bestimmen. Lösungspunkte von (1.1) für benachbarte Werte von λ lassen sich in diesem Fall mit auf dem Newtonverfahren aufbauenden Kurvenverfolgungsverfahren bestimmen. Sollte die Jacobimatrix $\partial_x F$ dagegen im Lösungspunkt singulär sein, ist dies nicht so einfach möglich. Solche Lösungspunkte nennt man singuläre Punkte.

Seit vielen Jahren wurden erweiterte Systeme zur Berechnung von verschiedenen singulären Punkten vorgeschlagen. Eine kurzer historischer Abriß steht am Anfang des Abschnitts 2.2. Diese erweiterten Systeme besitzen den singulären Punkt als reguläre Lösung. Deshalb kann auf die erweiterten Systeme das Newtonverfahren unter Ausnutzung der überlinearen Konvergenz zur Berechnung singulärer Punkte angewendet werden.

In dieser Arbeit soll es um eine geschlossene Darstellung der Vorgehensweise zur Berechnung singulärer Punkte mittels Klassen erweiterter Systeme gehen. Diese Vorgehensweise wird auf viele aus der Literatur bekannten singulären Punkte angewendet. Dabei zeigt sich, daß viele aus der Literatur bekannten Vorgehensweisen zur Berechnung singulärer Punkte sich unmittelbar aus dieser allgemeinen Vorgehensweise ergeben. Weiterhin wird diese allgemeine Vorgehensweise untersucht, insbesondere ihre Korrektheit nachgewiesen. Daraus

ergeben sich für weitere singuläre Punkte Berechnungsmöglichkeiten bzw. neue Berechnungsmöglichkeiten für viele singuläre Punkte. Diese werden in dieser Arbeit aufgeführt. In dieser Arbeit werden nur aus der Literatur bekannte singuläre Punkte behandelt. An den entsprechenden Stellen wird jedoch darauf hingewiesen, wie die Anwendung der allgemeinen Vorgehensweise auf weitere Arten singulärer Punkte aussehen könnte. Zur Erhaltung der Allgemeinheit werden in dieser Arbeit u.a. sehr allgemeine Werte von m, n, p und k und $q = m - \operatorname{rank}(\partial_x F)$, dem Zeilenrangabfall der Jacobimatrix im singulären Punkt, zugelassen. Bei allen konkret betrachteten singulären Punkten gilt jedoch $m \in \{n; n-1\}$, $q \in \{1; 2\}$ und $p \in \{0; 1\}$.

Diese allgemeine Vorgehensweise wird zusammen mit dem Aufbau dieser Arbeit erläutert.

Im Kapitel 2 werden die benötigten erweiterten Systeme hergeleitet. Dabei werden im Abschnitt 2.1 zuerst die Funktion F, der singuläre Punkt $(x^*, \lambda^*, \alpha^*)$, erweiterte und minimal erweiterte Systeme und einige später benötigte Hilfsgrößen und -matrizen definiert. Diese erweiterten Systeme enthalten neben der Ausgangsgleichung (1.1) noch weitere den singulären Punkt charakterisierende Gleichungen und bei Bedarf weitere Parameter. Wie zu Beginn des Abschnitts 2.2 erwähnt wird, schlugen bereits Jepson/Spence [JS84] vor, diese charakterisierenden Gleichungen aus den definierenden Gleichungen des singulären Punktes zu gewinnen. Weitere Autoren wie z.B. Govaerts [Gov97a] führten dies für eine Reihe singulärer Punkte und für spezielle Vorgehensweisen aus. In diesen definierenden Gleichungen kommen Ableitungen der Ljapunov-Schmidt-reduzierten Funktion vor. Deshalb werden zuerst verschiedene Ljapunov-Schmidt-reduzierte Funktionen aus der Literatur beschrieben und untersucht. Bei ihrer Definition werden verschiedene Regularitätsannahmen gemacht. In dieser Arbeit wird nachgewiesen, daß sie alle gleichzeitig erfüllt bzw. nicht erfüllt also äquivalent sind. In Verallgemeinerung dieser Vorgehensweisen wird eine verallgemeinerte Ljapunov-Schmidt-reduzierte Funktion auf zwei verschiedene Arten eingeführt und die Gleichheit dieser Definitionen gezeigt. Eine dieser Einführungsarten wird daraufhin weiter verallgemeinert. Es wird gezeigt, daß alle verallgemeinerten Ljapunov-Schmidt-Reduktionen einer gegebenen Funktion in der Umgebung eines gegebenen singulären Punktes bezüglich einer in diesem Abschnitt definierten Äquivalenzrelation äquivalent sind. Diese Äquivalenzrelation ist eine Verallgemeinerung der Kontaktäquivalenz aus [GS85], die dort für $m = n$ und $p = k = 0$ definiert wurde, auf die in dieser Arbeit betrachteten Dimensionen m, p, k und n. Der Nachweis dieser Äquivalenz erfolgt analog zum Nachweis aus [GS85]. Mit diesen verallgemeinerten Ljapunov-Schmidt-reduzierten Funktionen können nun charakterisierende Gleichungen konstruiert werden, die im erweiterten System verwendet werden können. Die benötigten Ableitungen können aus den Ableitungen des geränderten Systems, mit dem die verallgemeinerte Ljapunov-Schmidt-reduzierte Funktion definiert wird, ohne große Probleme bestimmt werden. Es sind lediglich lineare Gleichungssysteme mit einer nur vom aktuellen Punkt und nicht von der benötigten Ableitung abhängige Systemmatrix zu lösen. Diese Gleichungssysteme enthalten nur Ableitungen von F im aktuellen Punkt und bekannte Matrizen. Wenn das erweiterte System mit dem Newtonverfahren gelöst wird, wird neben dem Wert der charakterisierenden Funktionen, der linken Seite der charakterisierenden Gleichungen, im aktuellen Punkt auch noch die Jacobimatrix des erweiterten Systems benötigt. Speziell werden gewisse Ableitungen der charakterisierenden Funktionen gebraucht. Bei Verwendung minimal erweiterter Systeme und der verallgemeinerten Ljapunov-Schmidt-reduzierte Funktion sind diese jedoch im Regelfall nicht ganz einfach ermittelbar. Deshalb werden im minimal erweiterten System die Ableitungen einer anderen reduzierten Funktion

verwendet. Diese wird im Unterabschnitt 2.2.5 eingeführt. Wenn diese reduzierte Funktion in die charakterisierenden Gleichungen der entsprechenden singulären Punkte eingesetzt wird, ergeben sich gerade die in [PS81], [Pön87], [Pön90] und [PSS99] verwendeten charakterisierenden Gleichungen. Um nachweisen zu können, daß diese reduzierte Funktion zur Charakterisierung singulärer Punkte geeignet ist, wird im Unterabschnitt 2.2.6 nachgewiesen, daß diese reduzierte Funktion zur Ljapunov-Schmidt-reduzierten Funktionen im Sinne einer im Unterabschnitt 2.2.4 definierten Äquivalenzrelation äquivalent ist. Dabei wurde diese Äquivalenzrelation gerade so eingeführt, daß bei allen in dieser Arbeit betrachteten Arten singulärer Punkte äquivalente Funktionen die definierenden Gleichungen und die Nichtentartungsbedingungen, die zusammen den Typ des singulären Punktes eindeutig bestimmen, im singulären Punkt gleichzeitig erfüllen oder gleichzeitig nicht erfüllen. Damit läßt sich die im Unterabschnitt 2.2.5 eingeführte reduzierte Funktion zur Charakterisierung und Klassifizierung und damit in minimal erweiterten Systemen zur Berechnung singulärer Punkte verwenden. Diese Klassifikation erfolgt im Abschnitt 2.3. Daneben stehen einige Bemerkungen zu vorhandenen Klassifikationen singulärer Punkte und zu möglichen Erweiterungen der in dieser Arbeit vorgestellten Klassifikation auf weitere singuläre Punkte. Für alle in diesem Abschnitt klassifizierten singulären Punkte werden im Unterabschnitt 2.4.2 minimal erweiterte Systeme angegeben. Vorher erfolgen zu Beginn des Abschnitts 2.4 einige grundlegende Bemerkungen zur Wahl der charakterisierenden Gleichungen im erweiterten System. Insbesondere wird gezeigt, wie mittels der den singulären Punkt charakterisierenden definierenden Gleichungen und Nichtentartungsbedingungen nachgewiesen werden kann, daß der singuläre Punkt eine reguläre Lösung des erweiterten Systems ist. Zum Schluß werden im Unterabschnitt 2.4.3 viele erweiterte Systeme aus der Literatur aufgeführt, die sich unmittelbar aus der in dieser Arbeit beschriebenen Vorgehensweise ergeben bzw. in enger Beziehung mit ihr stehen. Damit ist gezeigt, daß die mit der verallgemeinerten Ljapunov-Schmidt-reduzierten Funktion bzw. der im Unterabschnitt 2.2.5 eingeführten reduzierten Funktion konstruierten erweiterten Systeme tatsächlich Verallgemeinerungen vieler in der Literatur vorgeschlagenen erweiterten Systeme sind. Kapitel 2 ist eine Überarbeitung und teilweise Erweiterung von [SPJ99]. Diese Arbeit geht inhaltlich im Wesentlichen auf den Autor der vorliegenden Arbeit zurück, während die konkrete englische Formulierung von Herrn Dr. Pönisch erfolgte. Im Abschnitt 2.1 und dem Unterabschnitt 2.4.3 wurden auch Texte aus [Sch98] und [SW99] verwendet.

Im Abschnitt 2.2 wurde nachgewiesen, daß alle mit einem geränderten System gleicher Dimension definierten verallgemeinerten Ljapunov-Schmidt-reduzierten Funktionen zueinander verallgemeinert kontaktäquivalent sind. Deshalb können die Ränderungen in diesem geränderten System bei Einhaltung der Dimension und der Regularität der Jacobimatrix dieses geränderten Systems geeignet gewählt werden. Eine sinnvolle Forderung ist sicher, daß diese Jacobimatrix möglichst kleine Konditionszahl besitzt. Schließlich ist sie die Systemmatrix aller für einen Punkt (x, λ, α) zu lösenden linearen Gleichungssysteme zur Berechnung der im Newtonverfahren benötigten charakterisierenden Funktionen und deren Ableitungen. Bei Verwendung des zweistufigen Newtonverfahrens aus Kapitel 4 ist neben den linearen Systemen mit dieser Systemmatrix lediglich ein kleines System mit einer anderen Systemmatrix zu lösen. Diese andere Systemmatrix besteht aus Richtungsableitungen der charakterisierenden Funktion im aktuellen Punkt. Deshalb werden im Kapitel 3 solche die Konditionszahl minimierende Ränderungen ermittelt. Am Anfang dieses Kapitels stehen einige Literaturverweise auf Arbeiten, die sich mit der Ränderung von Matrizen bei der Berechnung singulärer Punkte befassen. Im Abschnitt 3.1 werden notwendige und

hinreichende Bedingungen für die Minimalität der Konditionszahl angegeben. Dies ist eine Verallgemeinerung von Ergebnissen aus [EHM95]. Konkret kann die Ränderung mit gewissen Vielfachen der Singulärvektoren zu den kleinsten Singulärwerten erfolgen. Diese werden im Abschnitt 3.2 näherungsweise bestimmt. Dafür wird die inverse Teilraumiteration für Singulärwerte verwendet. Eine Übersicht über einige Arbeiten zur inversen Vektoriteration für Singulärwerte und zur inversen Teilraumiteration für Eigenwerte und deren Vorläuferalgorithmen wird zum Beginn des Abschnitts 3.2 angegeben. Außerdem werden andere sich aus der Literatur ergebende Verfahren zur Bestimmung der Singulärvektoren zu den kleinsten Singulärwerten erwähnt. Schließlich wird noch eine Inhaltsübersicht zu diesem Abschnitt gegeben. Speziell wird in diesem Abschnitt ein Verfahren mit den gewünschten Eigenschaften entwickelt. Zahlreiche numerische Beispiele zeigen die Anwendbarkeit dieses Verfahrens. Da die inverse Teilraumiteration nur eine linear konvergente Folge liefert, wird zum Schluß von Kapitel 3 noch ein Ausblick auf die Anwendung des Newtonverfahrens zur Bestimmung der gesuchten Singulärvektoren und -werte gegeben. Diese basieren auf den entsprechenden Überlegungen für Eigenwerte aus [SL97] und [LST98].

Im Kapitel 4 wird schließlich untersucht, wie sich die minimal erweiterten Systeme aus Kapitel 2 mit Hilfe des Newtonverfahrens lösen lassen. Um nicht alle Ableitungen der charakterisierenden Funktionen berechnen zu müssen, wird ein zweistufiges Newtonverfahren eingeführt. Dieses geht auf [PS81] zurück. Eine Übersicht über die für dieses Kapitel verwendete Literatur und den Aufbau dieses Kapitels stehen an seinem Anfang. Insbesondere ist dieses Kapitel eine Überarbeitung von [SW99]. Analog zum Kapitel 2 geht der Inhalt im Wesentlichen auf den Autor der vorliegenden Arbeit zurück. Bei den Formulierungen hat Frau Dr. Walther wesentlich geholfen. Teilweise wurden auch Formulierungen aus [PSS99] übersetzt, die von Herrn Dr. Pönisch geprägt wurden. In diesem Abschnitt wird gezeigt, daß sich das allgemeine minimal erweiterte System aus Kapitel 2, speziell aus dem Abschnitt 2.4, mit diesem zweistufigen Newtonverfahren unter den üblichen Bedingungen an das Newtonverfahren lösen läßt. Dies gilt nicht nur für die im Unterabschnitt 2.4.2 definierten erweiterten Systeme. Alle benötigten Ableitungen lassen sich mit Hilfe der Automatischen Differentiation bestimmen. Dem Autor der vorliegenden Arbeit ist außer den Arbeiten, an denen er selbst beteiligt war, keine Arbeit bekannt, in der die Verwendung der Automatischen Differentiation zur Berechnung singulärer Punkte beschrieben wird.

Die Ergebnisse der vorliegende Arbeit können sicher auch bei anderen in der Literatur behandelten Problemen verwendet werden.

So wurden in neuerer Zeit singuläre Punkte in nichtlinearen Gleichungssystemen mit Symmetrieeigenschaften unter Ausnutzung dieser Symmetrien bestimmt. Ebenfalls wurde vorgeschlagen, die erweiterten Systeme für singuläre Punkte zur Parameterschätzung zu nutzen. Um diese beiden Probleme geht es jedoch in der vorliegenden Arbeit nicht. Jedoch lassen sich sicherlich die in dieser Arbeit vorgeschlagenen erweiterten Systeme auch bei solchen Problemen verwenden.

Weiterhin wurden in verschiedenen Arbeiten die dort verwendeten erweiterten Systeme zur Kurvenverfolgung von Kurven solcher singulärer Punkte verwendet. Dabei wurden ausgehend von einem Lösungspunkt von (1.1) weitere Lösungspunkte von (1.1) mittels eines Kurvenverfolgungsalgorithmus bei einem variablen Parameter berechnet. Dabei wurde beispielsweise mittels Vorzeichenbetrachtungen gewisser Größen untersucht, ob zwischen zwei berechneten Lösungspunkten ein singulärer Punkt liegt. Wenn ein solcher vorhanden ist, wurde er mittels eines geeigneten erweiterten Systems berechnet. Ausgehend von diesem singulären Punkt wurde ein Kurvenverfolgungsalgorithmus zur Berechnung weiterer derar-

tiger singulärer Punkte verwendet. Dabei wurde ein weiterer Parameter als variabel aufgefaßt. Die beim Kurvenverfolgungsverfahren verwendeten erweiterten Systeme ergeben sich einfach aus dem erweiterten System für den entsprechenden singulären Punkt. Auf dieser Kurve singulärer Punkte lassen sich analog singuläre Punkte höherer Kodimension feststellen. Diese können dann wieder der Ausgangspunkt für die Berechnung weiterer gleichartiger singulärer Punkte mittels Kurvenverfolgung sein. Die Kodimension beschreibt beispielsweise, wieviel Parameter α vorhanden sein müssen, d.h., wie groß k mindestens sein muß, damit dieser singulärer Punkt gegenüber Störungen von F stabil ist. Stabilität bedeutet in diesem Fall, daß alle Funktionen, die von F normmäßig nur wenig abweichen, in der Umgebung des singulären Punktes von F ebenfalls einen gleichartigen singulären Punkt besitzen. Je größer die Kodimension ist, desto mehr Parameter werden für die Stabilität benötigt und desto seltener treten somit solche singulären Punkte auf. Somit lassen sich durch die fortgesetzte Anwendung des beschriebenen Verfahrens singuläre Punkte größerer Kodimension bestimmen. Ausgehend von diesen Punkten können auch Kurven singulärer Punkte niedrigerer Kodimension bestimmt werden. Dabei werden bestimmte Parameter als fest betrachtet. Auch darum geht es in dieser Arbeit nicht. Jedoch werden auch in diesen Fällen geeignete erweiterte Systeme für die gesuchten singulären Punkte benötigt. So können z.B. die in dieser Arbeit vorgeschlagenen erweiterten Systeme dafür verwendet werden.

Kapitel 2

Erweiterte Systeme für singuläre Punkte

2.1 Definitionen

Die mathematische Beschreibung vieler praktischer Probleme führt zu nichtlinearen Gleichungssystemen der Form

$$F(x, \lambda, \alpha) = 0 \tag{2.1}$$

mit

$$F : \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^k \to \mathbb{R}^m : (x, \lambda, \alpha) \mapsto F(x, \lambda, \alpha) \tag{2.2}$$

und

$$n + p \geq m > 0\,, \quad n > 0\,, \quad p \geq 0 \text{ und } k \geq 0.$$

Der Vektor der Zustandsvariablen wird mit x bezeichnet, λ sei ein ausgezeichneter Parameter bzw. der Vektor der ausgezeichneten Parameter, auch Bifurkations- bzw. Verzweigungsparameter genannt. Die übrigen Parameter seien mit α bezeichnet. Diese Aufspaltung der Argumente ergibt sich häufig direkt aus der Problemstellung. Falls ein Parameter einen festen Wert hat, wird er nicht (λ, α) zugeordnet. Hier und im Folgenden kennzeichnet die Dimension 0, daß der entsprechende Parameter nicht existiert. Außerdem wird $(x^T, \lambda^T, \alpha^T)^T$ mit (x, λ, α) identifiziert.

Definition 2.1. Ein Punkt $(x^*, \lambda^*, \alpha^*)$ heißt singulärer Punkt von F, falls

$$F(x^*, \lambda^*, \alpha^*) = 0 \quad \text{und} \quad \operatorname{rank} \partial_x F(x^*, \lambda^*, \alpha^*) = m - q$$

mit $\max\{1, m - n + 1\} \leq q \leq m$ gilt.

Speziell ist im praktisch kaum vorkommenden Fall $q = m$: $\partial_x F(x^*, \lambda^*, \alpha^*) = 0 \in \mathbb{R}^{m \times n}$.
Im Folgenden sei $(x^*, \lambda^*, \alpha^*)$ ein singulärer Punkt, für den

$$\operatorname{rank} \partial F(x^*, \lambda^*, \alpha^*) = m - l \geq m - q \tag{2.3}$$

gilt. Die Funktion F wird in einer Umgebung von $(x^*, \lambda^*, \alpha^*)$ als hinreichend glatt vorausgesetzt; alle weiteren Betrachtungen sollen nur in dieser Umgebung erfolgen.

Wegen $\operatorname{rank} \partial_x F(x^*, \lambda^*, \alpha^*) = m - q$ und $\partial_x F(x^*, \lambda^*, \alpha^*) \in \mathbb{R}^{m \times n}$ gilt

$$n - m + q = \dim \ker \partial_x F(x^*, \lambda^*, \alpha^*), \quad q = \dim \ker \partial_x F(x^*, \lambda^*, \alpha^*)^T,$$

wobei $\ker A := \{x : A\,x = 0\}$ der Nullraum der Matrix A und $\dim \ker A$ die Dimension des Nullraumes von A ist. Weiterhin sei $\operatorname{im} A := \{y : y = A\,x, \ x \text{ beliebig}\}$ der Bildraum von A. Dieser wird von den Spalten von A aufgespannt.

Damit existieren Matrizen

$$U_r \in \mathbb{R}^{m \times (m-q)}, \quad U_s \in \mathbb{R}^{m \times q}, \quad V_r \in \mathbb{R}^{n \times (m-q)} \quad \text{und} \quad V_s \in \mathbb{R}^{n \times (n-m+q)} \tag{2.4}$$

so, daß $V := \begin{bmatrix} V_r & V_s \end{bmatrix}$, $U := \begin{bmatrix} U_r & U_s \end{bmatrix}$ orthogonale Matrizen sind und

$$\begin{aligned}
&\operatorname{im} U_s = \ker \partial_x F(x^*, \lambda^*, \alpha^*)^T, \quad \text{d.h.,} \quad U_s^T\, \partial_x F(x^*, \lambda^*, \alpha^*) = 0 \in \mathbb{R}^{q \times n}, \\
&\operatorname{im} U_r = \operatorname{im} \partial_x F(x^*, \lambda^*, \alpha^*), \\
&\operatorname{im} V_s = \ker \partial_x F(x^*, \lambda^*, \alpha^*), \quad \text{d.h.,} \quad \partial_x F(x^*, \lambda^*, \alpha^*)\, V_s = 0 \in \mathbb{R}^{m \times (n-m+q)}, \\
&\operatorname{im} V_r = \operatorname{im} \partial_x F(x^*, \lambda^*, \alpha^*)^T
\end{aligned} \tag{2.5}$$

gilt, d.h., die Matrizen (2.5) bilden orthogonale Basen des Nullraums, des Bildraums und deren orthogonaler Komplemente von $\partial_x F$ im singulären Punkt. Diese Matrizen könnten beispielsweise mittels Singulärwertzerlegung von $\partial_x F(x^*, \lambda^*, \alpha^*)$ bestimmt werden. Im Fall $q = m$ ist $m - q = 0$. Dies bedeutet analog zu den Aussagen bei den Parametern, daß in diesem Fall U_r und V_r nicht existieren.

In dieser Arbeit werden einige Eigenschaften von Projektoren genutzt. Diese werden im Folgenden zusammengestellt.

Definition 2.2. Ein Projektor ist ein linearer Operator

$$P : \mathbb{R}^n \mapsto \operatorname{im} P \subset \mathbb{R}^n \qquad \text{in Richtung von } \ker P \qquad \text{mit } P\,P = P.$$

Zur Beschreibung eines linearen Operators P die Matrix wird $P \in \mathbb{R}^{n \times n}$ verwendet. Für einen orthogonalen Projektor gilt $P = P^T$.

Es sei nun $P \in \mathbb{R}^{n \times n}$ ein Projektor, die Spalten von P_r seien eine Basis von $\operatorname{im} P$ und die Spalten von P_s seien eine Basis von $\ker P$. Dann läßt sich P gemäß

$$P = \begin{bmatrix} P_r & 0 \end{bmatrix} \begin{bmatrix} P_r & P_s \end{bmatrix}^{-1} \tag{2.6}$$

darstellen. Offensichtlich ist nämlich (2.6) ein Projektor,

$$\ker P = \operatorname{im} P_s \quad \text{und} \quad \operatorname{im} P = \operatorname{im} P_r.$$

Wie leicht einzusehen ist, hängt (2.6) nicht von der speziellen Wahl der Basen von $\operatorname{im} P$ und $\ker P$ ab. Bei der Wahl orthogonaler Basen ist ein orthogonaler Projektor P gemäß

$$P = P_r\, P_r^T$$

darstellbar.

Zur Vereinfachung werden die Bezeichnungen

$$M_{i,j} := \begin{bmatrix} I_j \\ 0 \end{bmatrix} \in \mathbb{R}^{i \times j}, \;\; i \geq j \quad \text{und} \quad N_{i,j} := \begin{bmatrix} 0 \\ I_j \end{bmatrix} \in \mathbb{R}^{i \times j}, \;\; i \geq j$$

eingeführt, wobei $I_j \in \mathbb{R}^{j \times j}$ die Einheitsmatrix ist.

Da das Ziel dieser Arbeit in der Berechnung singulärer Punkte nach Definition 2.1 der Funktion F (2.2) mit Hilfe des Newtonverfahrens unter Ausnutzung der überlinearen Konvergenz des Newtonverfahrens besteht, werden erweiterte Systeme eingeführt. Diese genügen folgender Definition:

Definition 2.3. Ein System

$$G(y) = 0, \quad G : \mathbb{R}^{m_{ges}} \to \mathbb{R}^{m_{ges}}, \quad m_{ges} \geq n + p + k, \tag{2.7}$$

heißt erweitertes System von $F(x, \lambda, \alpha) = 0$ bezüglich des singulären Punktes $(x^*, \lambda^*, \alpha^*)$, wenn ein Vektor $y^* \in \mathbb{R}^{m_{ges}}$ existiert, so daß

$$\begin{aligned} G(y^*) = 0, \qquad & M^T_{m_{ges}, n+p+k}\, y^* = (x^*, \lambda^*, \alpha^*), \\ \operatorname{rank} \partial G(y^*) = m_{ges}\,, \quad & M^T_{m_{ges}, m}\, \partial G(y^*)\, M_{m_{ges}, n+p+k} = \partial F(x^*, \lambda^*, \alpha^*) \end{aligned}$$

gilt. Wenn m_{ges} minimal ist, heißt (2.7) minimal erweitertes System.

Wenn in (2.3) $l = 0$ ist, ist $\partial F(x^*, \lambda^*, \alpha^*)$ zeilenregulär. Falls nun eine stetig differenzierbare Funktion

$$f : \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^k \to \mathbb{R}^{n+p+k-m} : (x, \lambda, \alpha) \mapsto f(x, \lambda, \alpha)$$

mit $f(x^*, \lambda^*, \alpha^*) = 0$ und $\begin{bmatrix} \partial F(x^*, \lambda^*, \alpha^*) \\ \partial f(x^*, \lambda^*, \alpha^*) \end{bmatrix}$ regulär existiert, so ist

$$G(y) := \begin{bmatrix} F(x, \lambda, \alpha) \\ f(x, \lambda, \alpha) \end{bmatrix} = 0 \in \mathbb{R}^{n+p+k}, \quad y := (x, \lambda, \alpha) \in \mathbb{R}^{n+p+k} \tag{2.8}$$

ein erweitertes System von $F(x, \lambda, \alpha) = 0$ bezüglich des singulären Punktes $(x^*, \lambda^*, \alpha^*)$ mit $y^* = (x^*, \lambda^*, \alpha^*)$. Da außerdem $m_{ges} = n + p + k$ ist, handelt es sich dabei sogar um ein minimal erweitertes System.

Wenn in (2.3) $l > 0$ ist, existiert eine Matrix

$$\bar{D} \in \mathbb{R}^{m \times l} \quad \text{mit} \quad \operatorname{rank} \begin{bmatrix} \partial F(x^*, \lambda^*, \alpha^*) & \bar{D} \end{bmatrix} = m. \tag{2.9}$$

Falls nun wieder eine stetig differenzierbare Funktion

$$f : \mathbb{R}^n \times \mathbb{R}^p \times \mathbb{R}^k \to \mathbb{R}^{n+p+k+l-m} : (x, \lambda, \alpha) \mapsto f(x, \lambda, \alpha)$$

mit $f(x^*, \lambda^*, \alpha^*) = 0$ und $\begin{bmatrix} \partial F(x^*, \lambda^*, \alpha^*) & \bar{D} \\ \partial f(x^*, \lambda^*, \alpha^*) & 0 \end{bmatrix}$ regulär existiert, so ist

$$G(y) := \begin{bmatrix} F(x, \lambda, \alpha) + \bar{D}\, \mu \\ f(x, \lambda, \alpha) \end{bmatrix} = 0 \in \mathbb{R}^{n+p+k+l}, \quad y := (x, \lambda, \alpha, \mu) \in \mathbb{R}^{n+p+k+l} \tag{2.10}$$

ein erweitertes System von $F(x, \lambda, \alpha) = 0$ bezüglich des singulären Punktes $(x^*, \lambda^*, \alpha^*)$ mit $y^* = (x^*, \lambda^*, \alpha^*, 0) \in \mathbb{R}^{n+p+k+l}$. Für ein erweitertes System $G(y) = 0$ von $F(x, \lambda, \alpha) = 0$ bezüglich des singulären Punktes $(x^*, \lambda^*, \alpha^*)$ ist wegen $\operatorname{rank} \partial G(y^*) = m_{ges}$ insbesondere $\operatorname{rank} M^T_{m_{ges},m} \partial G(y^*) = m$. Weiterhin ist

$$\operatorname{rank} M^T_{m_{ges},m} \partial G(y^*) M_{m_{ges},n+p+k} = \operatorname{rank} \partial F(x^*, \lambda^*, \alpha^*) = m - l,$$

weshalb $m_{ges} \geq n+p+k+l$ ist. Somit ist (2.10) in diesem Fall sogar ein minimal erweitertes System.

Damit beschränkt sich die Suche nach geeigneten erweiterten bzw. minimal erweiterten Systemen auf die Suche nach geeigneten Funktionen f für die jeweiligen singulären Punkte.

2.2 Reduzierte Funktionen

Erweiterte Systeme werden schon seit vielen Jahren zur Berechnung singulärer Punkte benutzt. Anfangs, z. B. von Seydel [Sey79] zur Berechnung einfacher Rückkehrpunkte, wurden dabei Systeme mit $m_{ges} \geq 2 * (n + p + k + l)$ vorgeschlagen. Die charakterisierenden Gleichungen $N^T_{m_{ges},m_{ges}-m} G(y) = 0$ in den erweiterten Systemen aus diesen Arbeiten charakterisieren den zu berechnenden singulären Punkt. In weiteren Arbeiten, z. B. von Pönisch/ Schwetlick [PS81], Jepson/Spence [JS84] und Griewank/Reddien [GR84], wurden eine bzw. wenige skalare Gleichungen zur Charakterisierung der singulären Punkte verwendet. Von diesen Gleichungen wurde gezeigt, daß sie genau dann erfüllt sind, wenn die charakterisierenden Gleichungen in den anfangs erwähnten großen erweiterten Systemen erfüllt sind. Auf diese Art konnte die Dimension der erweiterten Systeme drastisch reduziert werden. Bereits Jepson/Spence [JS84] schlugen vor, als charakterisierende Gleichungen die definierenden Gleichungen für den zu berechnenden singulären Punkt zu verwenden. Dies führt zu minimal erweiterten Systemen. Jedoch gab es damals bei den numerischen Berechnungen noch einige technische Probleme, die aber in den folgenden Jahren gelöst wurden. Daraus ergab sich eine einfache Vorgehensweise sowohl bei der Konstruktion und dem Nachweis der Eigenschaften erweiterter Systeme, als auch bei der numerischen Berechnung. Für eine spezielle Vorgehensweise wurde dies von Govaerts [Gov97a] vorgeführt. Parallel dazu wurden weitere erweiterte Systeme zur Berechnung spezieller singulärer Punkte eingeführt; für eine umfassende Übersicht bis 1989 kann auf Kunkel [Kun91, Abschnitt 9.1, S. 120ff.] verwiesen werden. Dabei stellt sich heraus, daß viele der vorgeschlagenen charakterisierenden Gleichungen gerade den definierenden Gleichungen der zu berechnenden singulären Punkte entsprachen bzw. mit diesen in einem engen Zusammenhang standen. Daraus ergeben sich zusätzlich zu der von Govaerts [Gov97a] vorgeschlagene Vorgehensweise weitere mögliche Vorgehensweisen bei der Konstruktion und der Nutzung reduzierter Funktionen in erweiterten Systemen. Die entsprechenden reduzierten Funktionen sollen in diesem Abschnitt hergeleitet werden. Dabei werden die bekannten Vorgehensweisen aus der Literatur systematisiert, erweitert bzw. auf andere reduzierte Funktionen übertragen. Insbesondere wird für die von Pönisch, Schwetlick und Schnabel [PS81], [Pön87], [Pön90] und [PSS99] verwendeten charakterisierenden Gleichungen gezeigt, daß sie als definierenden Gleichungen der entsprechenden singulären Punkte bei Verwendung einer speziellen reduzierten Funktion interpretiert werden können. Es wird gezeigt, daß diese reduzierte Funktion für sehr viele bekannte singuläre Punkte die gleichen Eigenschaften wie die Ljapunov-Schmidt-reduzierte Funktion hat. Daraus ergeben sich für sehr viele singuläre Punkte neue erweiterte Systeme.

In diesem Abschnitt werden zuerst einige bekannte Ljapunov-Schmidt-Reduktionen bzw. verallgemeinerte Ljapunov-Schmidt-Reduktionen aus der Literatur dargestellt. Darauf aufbauend wird eine allgemeine Ljapunov-Schmidt-Reduktion auf zwei verschiedene Weisen eingeführt, deren Gleichheit gezeigt und Beziehungen zu den bekannten Ljapunov-Schmidt-Reduktionen erwähnt. Durch Übertragung von Äquivalenzrelationen aus der Literatur wird eine neue Äquivalenzrelation, eine verallgemeinerte Kontaktäquivalenz, eingeführt. Bezüglich dieser verallgemeinerten Kontaktäquivalenz wird gezeigt, daß alle allgemeinen Ljapunov-Schmidt-Reduktionen einer gegebenen Funktion in der Umgebung eines gegebenen singulären Punktes äquivalent sind. Schließlich wird die neue reduzierte Funktion eingeführt und gezeigt, daß sie bezüglich einer weiteren neueingeführten Äquivalenzrelation, der Ableitungsäquivalenz, äquivalent zu den allgemeinen Ljapunov-Schmidt-reduzierten Funktionen ist. Diese Ableitungsäquivalenz ermöglicht eine Klassifikation singulärer Punkte mittels dieser reduzierten Funktion analog zur Klassifikation mittels der Ljapunov-Schmidt-reduzierten Funktion, vergleiche Golubitsky/Schaeffer [GS85] und Govaerts [Gov95] und [Gov97a]. Diese Klassifikation erfolgt im folgenden Abschnitt 2.3. Dieser und der folgende Abschnitt 2.3 ist eine überarbeitete und teilweise erweiterte Version von [SPJ99].

In den folgenden Unterabschnitten werden verschiedene reduzierte Funktionen betrachtet. Diese werden implizit mittels nichtlinearer Systeme definiert, die aus (2.1) gewonnen werden. Um den Satz über die impliziten Funktionen anwenden zu können, sind gewisse Regularitätsbedingungen zu erfüllen. Im Satz 2.4 wird gezeigt, daß diese Regularitätsbedingungen von Golubitsky/Schaeffer [GS85], Beyn [Bey84], Jepson/Spence [JS84], Griewank/Reddien [GR89], Kunkel [Kun91] und Govaerts [Gov97a] äquivalent sind. Der Beweis ist trivial und wird deshalb weggelassen. Die Matrix W_2 geht auf Jepson/Spence [JS84] zurück, weshalb die dortige Bezeichnung verwendet wird. Die Bezeichnung der Matrizen Q_s und P_r sind dagegen Kombinationen aus den Bezeichnungen der Projektoren Q und P in [Kun91] und den Indizes aus (2.6) bzw. entsprechend (2.4) und (2.5).

Satz 2.4.
Es seien $q < m$, $T \in \mathbb{R}^{(n-m+q)\times n}$, $Q_s \in \mathbb{R}^{m\times q}$, $W_2 \in \mathbb{R}^{(m-q)\times m}$, $P_r \in \mathbb{R}^{n\times(m-q)}$ *und* $\delta \in \mathbb{R}^{(n-m+q)\times q}$, *wobei* $\ker T = \operatorname{im} P_r$ *und* $\ker W_2 = \operatorname{im} Q_s$. *Weiterhin seien* $\begin{bmatrix} V_r & V_s \end{bmatrix}$ *und* $\begin{bmatrix} U_r & U_s \end{bmatrix}$ *orthogonale Matrizen gemäß (2.4) und (2.5). Dann sind die folgenden Aussagen äquivalent.*

(a) $\begin{bmatrix} \partial_x F(x^*, \lambda^*, \alpha^*) & \pm Q_s \\ T & \pm\delta \end{bmatrix}$ *ist regulär.*

(b) $U_s^T Q_s$ *und* $T V_s$ *sind regulär.*

(c) $\begin{bmatrix} U_r & Q_s \end{bmatrix}$ *und* $\begin{bmatrix} P_r & V_s \end{bmatrix}$ *sind regulär.*

(d) $\begin{bmatrix} W_2^T & U_s \end{bmatrix}$ *und* $\begin{bmatrix} P_r & V_s \end{bmatrix}$ *sind regulär.*

(e) Die Spalten von W_2^T *bzw.* P_r *sind linear unabhängig und* $\operatorname{im} W_2^T \cap \operatorname{im} U_s = \{0\}$ *und* $\operatorname{im} P_r \cap \operatorname{im} V_s = \{0\}$.

(f) $W_2\, \partial_x F(x^*, \lambda^*, \alpha^*) P_r = W_2\, U_r\, U_r^T\, \partial_x F(x^*, \lambda^*, \alpha^*)\, V_r\, V_r^T\, P_r$ *ist regulär.*

(g) $Q\,\partial_v F(x^*,\lambda^*,\alpha^*)$ *ist ein regulärer Operator von* $\operatorname{im} P_r$ *auf* $\operatorname{im} Q$*, wobei* $v := Px$ *ist und* Q *und* P *Projektoren mit* $\ker Q = \operatorname{im} Q_s$ *und* $\operatorname{im} P = \operatorname{im} P_r$ *sind.*

Bemerkung 2.5.

(i) Aus $\ker T = \operatorname{im} P_r$ und $\ker W_2 = \operatorname{im} Q_s$ folgt, daß die Zeilen von T und von W_2 linear unabhängig und daß P_r und Q_s spaltenregulär sind. Falls die entsprechenden Matrizen in (a), (b), ... , (f) auftreten, kann dies auch aus den entsprechenden Aussagen geschlußfolgert werden.

(ii) Im Fall $q = m$ existieren U_r, W_2 und P_r nicht. Damit sind

$$\begin{bmatrix} U_r & Q_s \end{bmatrix} = Q_s\,, \quad \begin{bmatrix} P_r & V_s \end{bmatrix} = V_s \quad \text{und} \quad \begin{bmatrix} W_2^T & U_s \end{bmatrix} = U_s\,.$$

Dann ist (a) äquivalent zu (b), (d) ist laut Definition 2.5 erfüllt, während (e), (f) und (g) keine Aussagen liefern.

2.2.1 Reduzierte Funktionen in der Literatur

In diesem Unterabschnitt werden einige Ljapunov-Schmidt-Reduktionen und verallgemeinerte Ljapunov-Schmidt-Reduktionen aus der Literatur vorgestellt. Zwischen diesen gibt es Unterschiede in den verwendeten Projektoren, in den Dimensionen, in der Lage des singulären Punktes bezüglich des Koordinatenursprungs und in der Allgemeinheit der betrachteten Singularitäten.

Zur Beschreibung der Reduktionen werden U_r, U_s, V_r und V_s im Sinne von (2.4) und (2.5) benutzt. Die allgemeinen Matrizen

$$Q_r \in \mathbb{R}^{m\times(m-q)}, \quad Q_s \in \mathbb{R}^{m\times q}, \quad P_r \in \mathbb{R}^{n\times(m-q)} \quad \text{und} \quad P_s \in \mathbb{R}^{n\times(n-m+q)} \tag{2.11}$$

seien so gewählt, daß die entsprechend (2.6) definierten Matrizen P und Q Projektoren sind, vergleiche Kunkel [Kun91] für die Bezeichnung der Projektoren. Im Fall $q = m$ existieren Q_r und P_r nicht, Q_s und P_s sind regulär und P und Q enthalten nur Nullelemente.

Golubitsky/Schaeffer [GS85] betrachteten ein System mit $n = m$ und $p = 1$, wobei der singuläre Punkt (x^*,λ^*,α^*) im Ursprung $(0,0,0)$ liegt. Dabei untersuchten sie zuerst den Fall $q = 1$ und später den Fall $1 \le q < m$. Dabei betrachteten sie unter der Voraussetzung von Satz 2.4(c) das System

$$0 = \begin{bmatrix} U_r & 0 \end{bmatrix} \begin{bmatrix} U_r & Q_s \end{bmatrix}^{-1} F(v+w,\lambda,\alpha) \tag{2.12}$$

mit

$$v = \begin{bmatrix} 0 & V_s \end{bmatrix} \begin{bmatrix} P_r & V_s \end{bmatrix}^{-1} x =: V_s R_v\,\xi \quad \text{und} \quad w = \begin{bmatrix} P_r & 0 \end{bmatrix} \begin{bmatrix} P_r & V_s \end{bmatrix}^{-1} x\,.$$

Dabei ist $R_v \in \mathbb{R}^{(n-m+q)\times(n-m+q)}$ eine reguläre Matrix. Sie wurde eingeführt, da Golubitsky/Schaeffer [GS85] nicht notwendigerweise eine orthogonale Basis von $\ker \partial_x F(x^*,\lambda^*,\alpha^*)$ benutzten. Analog wird eine reguläre Matrix $R_u \in \mathbb{R}^{q\times q}$ eingeführt, so daß die Spalten von $U_s R_u$ die verwendeten Basisvektoren von $\ker \partial_x F(x^*,\lambda^*,\alpha^*)^T$ sind.

Lemma 2.6.
Unter den Voraussetzungen vom Satz 2.4 im Fall $(x^*, \lambda^*, \alpha^*) = (0,0,0)$ *ist die Voraussetzung aus Satz 2.4(c) äquivalent zu Satz 2.4(f). Dies läßt sich auch als*

$$M^T_{m,m-q} \begin{bmatrix} U_r & Q_s \end{bmatrix}^{-1} \partial_x F(0,0,0)\, P_r \quad \textit{ist regulär}$$

schreiben, wobei $W_2 = M^T_{m,m-q} \begin{bmatrix} U_r & Q_s \end{bmatrix}^{-1}$ *ist, d. h.,* $W_2 U_r = I_{m-q}$.

Bemerkung 2.7.

(i) $U_r\, M^T_{m,m-q} = \begin{bmatrix} U_r & 0 \end{bmatrix}$.

(ii) $M^T_{m,m-q} \begin{bmatrix} U_r & Q_s \end{bmatrix}^{-1} \partial_x F(0,0,0)\, P_r$ ist der w-Anteil der Ableitung von (2.12) nach x.

(iii) $\xi = R_v^{-1} N^T_{n,n-m+q} \begin{bmatrix} P_r & V_s \end{bmatrix}^{-1} x$ ist eine Parametrisierung des v-Anteils von x.

Wegen des Satzes über die impliziten Funktionen und wegen Lemma 2.6 existiert in einer Umgebung von $(0,0,0) \in \operatorname{im} V_s \times \mathbb{R}^1 \times \mathbb{R}^k$ eine lokale Funktion

$$w : \operatorname{im} V_s \times \mathbb{R}^1 \times \mathbb{R}^k \to \operatorname{im} P_r : (V_s R_v\, \xi, \lambda, \alpha) \mapsto w(V_s R_v\, \xi, \lambda, \alpha)$$

mit $w(0,0,0) = 0$ und

$$0 \equiv \begin{bmatrix} U_r & 0 \end{bmatrix} \begin{bmatrix} U_r & Q_s \end{bmatrix}^{-1} F(V_s R_v\, \xi + w(V_s R_v\, \xi, \lambda, \alpha), \lambda, \alpha)\,. \tag{2.13}$$

Mit dieser Funktion w wird nun die Ljapunov-Schmidt-reduzierte Funktion g definiert.

Definition 2.8. Eine lokal definierte Funktion

$$g : \mathbb{R}^q \times \mathbb{R}^1 \times \mathbb{R}^k \to \mathbb{R}^q : (\xi, \lambda, \alpha) \mapsto g(\xi, \lambda, \alpha) := R_u^T U_s^T F(V_s R_v\, \xi + w(V_s R_v\, \xi, \lambda, \alpha), \lambda, \alpha)$$

heißt Ljapunov-Schmidt-reduzierte Funktion, wobei w implizit durch (2.13) definiert ist.

Die Definition 2.8 ist allgemeiner als die von Kunkel [Kun91], bei der der Parameter α nicht auftritt, d. h. $k = 0$, und orthogonale Projektoren, d. h. $Q_s = U_s, P_r = V_r$, und orthonormale Basen, d. h. $R_v = I_{n-m+q}$ und $R_u = I_q$, genutzt werden. Jedoch betrachtet Kunkel auch die Fälle $n \geq m$ und $p \geq 0$. Analog zu (2.13) erhält man damit lokal

$$0 \equiv U_r\, U_r^T F(V_s\, \xi + w(V_s\, \xi, \lambda), \lambda) \tag{2.14}$$

mit

$$v := V_s\, \xi := V_s\, V_s^T\, x \quad \text{und} \quad w := V_r\, V_r^T\, x\,.$$

Somit ergibt sich analog zu Definition 2.8 die folgende Definition.

Definition 2.9. Eine lokal definierte Funktion

$$g : \mathbb{R}^{n-m+q} \times \mathbb{R}^p \to \mathbb{R}^q : (\xi, \lambda) \mapsto g(\xi, \lambda) := U_s^T F(V_s\, \xi + w(V_s\, \xi, \lambda), \lambda)$$

heißt Ljapunov-Schmidt-reduzierte Funktion, wobei w implizit durch (2.14) definiert ist.

Beyn [Bey84] betrachtete den Fall $n \geq m$ und $p = 0$, d. h., λ tritt nicht auf. Da U_r, V_s und U_s i. allg. nicht von vornherein bekannt sind, verwendete er sie nicht in seiner Definition. Weiterhin setzte er nicht voraus, daß der singuläre Punkt im Koordinatenursprung $(0,0)$ liegt. Analog zu (2.12) betrachtete Beyn das System

$$0 = \begin{bmatrix} Q_r & 0 \end{bmatrix} \begin{bmatrix} Q_r & Q_s \end{bmatrix}^{-1} F(v + w, \alpha) \tag{2.15}$$

mit

$$v = \begin{bmatrix} P_r & 0 \end{bmatrix} \begin{bmatrix} P_r & P_s \end{bmatrix}^{-1} x \quad \text{und} \quad w = \begin{bmatrix} 0 & P_s \end{bmatrix} \begin{bmatrix} P_r & P_s \end{bmatrix}^{-1} x \,,$$

wobei $\begin{bmatrix} Q_r & Q_s \end{bmatrix}$ und $\begin{bmatrix} P_r & P_s \end{bmatrix}$ regulär seien. Weiterhin seien die Projektoren $Q = \begin{bmatrix} Q_r & 0 \end{bmatrix} \begin{bmatrix} Q_r & Q_s \end{bmatrix}^{-1}$ und $\begin{bmatrix} P_r & 0 \end{bmatrix} \begin{bmatrix} P_r & P_s \end{bmatrix}^{-1}$ so gewählt, daß Satz 2.4(g) im Fall $p = 0$ erfüllt ist, d. h., $Q\, \partial_v F$ sei ein regulärer Operator von $\operatorname{im} P_r$ auf $\operatorname{im} Q_r$ im singulären Punkt (x^*, α^*). Die Anwendung des Satzes über die impliziten Funktionen auf (2.15) ergibt die Existenz der Funktion $v : \operatorname{im} P_s \times \mathbb{R}^k \to \operatorname{im} P_r : (w, \alpha) \mapsto v(w, \alpha)$ mit

$$v(\begin{bmatrix} 0 & P_s \end{bmatrix} \begin{bmatrix} P_r & P_s \end{bmatrix}^{-1} x^*, \alpha^*) = \begin{bmatrix} P_r & 0 \end{bmatrix} \begin{bmatrix} P_r & P_s \end{bmatrix}^{-1} x^*$$

und

$$0 \equiv \begin{bmatrix} Q_r & 0 \end{bmatrix} \begin{bmatrix} Q_r & Q_s \end{bmatrix}^{-1} F(v(w, \alpha) + w, \alpha) \tag{2.16}$$

in einer Umgebung von $(\begin{bmatrix} 0 & P_s \end{bmatrix} \begin{bmatrix} P_r & P_s \end{bmatrix}^{-1} x^*, \alpha^*)$. Im Gegensatz zu Golubitsky/Schaeffer und zu Kunkel definierte Beyn eine verallgemeinerte Ljapunov-Schmidt-Reduktion.

Definition 2.10. Eine lokal definierte Funktion

$$\boldsymbol{g} : \operatorname{im} P_s \times \mathbb{R}^k \to \operatorname{im} Q_s : (w, \alpha) \mapsto \boldsymbol{g}(w, \alpha) := \begin{bmatrix} 0 & Q_s \end{bmatrix} \begin{bmatrix} Q_r & Q_s \end{bmatrix}^{-1} F(v(w, \alpha) + w, \alpha)$$

heißt verallgemeinerte Ljapunov-Schmidt-reduzierte Funktion, wobei v implizit durch (2.16) definiert ist.

Jepson/Spence [JS84] betrachteten wie Golubitsky/Schaeffer [GS85] ebenfalls den Fall $n = m$, $p = 1$, jedoch nur $q = 1$. Anstatt U_r, V_s und U_s verwendeten sie

$$v_1 \in \mathbb{R}^n, \quad V_2 \in \mathbb{R}^{n \times (n-1)}, \quad w_1 \in \mathbb{R}^m \quad \text{und} \quad W_2 \in \mathbb{R}^{(m-1) \times m} \tag{2.17}$$

mit den Eigenschaften

$$V_s \notin \operatorname{im} V_2 \,, \quad U_s \notin \operatorname{im} W_2^T, \quad \begin{bmatrix} v_1 & V_2 \end{bmatrix} \quad \text{und} \quad \begin{bmatrix} w_1 & W_2^T \end{bmatrix} \quad \text{sind regulär,} \tag{2.18}$$

woraus Satz 2.4(e) im Fall $P_r = V_2$ folgt. Die ebenfalls von Jepson/Spence [JS84] geforderten Beziehungen $w_1^T U_s \neq 0$, $v_1^T V_s \neq 0$ erweisen sich als unnötig. Das (2.12) entsprechende System lautet

$$0 = W_2 F(x^* + V_2 \eta + v_1 \xi, \lambda, \alpha) \quad \text{mit} \quad \begin{bmatrix} \eta \\ \xi \end{bmatrix} = \begin{bmatrix} V_2 & v_1 \end{bmatrix}^{-1} (x - x^*) \,. \tag{2.19}$$

Bemerkung 2.11.

- Golubitsky/Schaeffer verwendeten für v_1 und w_1 die speziellen Vektoren $v_1 = V_s R_v$ und $w_1 = U_s R_u$.
- Hier ist der erste Fall, bei dem der nichttriviale Wert von x^* in der Gleichung vorkommt.
- $x = x^* + V_2\eta + v_1\xi \quad \Longleftrightarrow \quad \begin{bmatrix} \eta \\ \xi \end{bmatrix} = \begin{bmatrix} V_2 & v_1 \end{bmatrix}^{-1} (x - x^*)\,.$

Lemma 2.12.
Unter den Voraussetzungen vom Satz 2.4(e) im Fall $m = n$, $p = q = 1$ und $P_r = V_2$ gilt Satz 2.4(f), d. h.,

$$W_2\,\partial_\eta F(x^*, \lambda^*, \alpha^*) = W_2\,\partial_x F(x^*, \lambda^*, \alpha^*) V_2 \quad \textit{ist regulär.}$$

Nach dem Satz über die impliziten Funktionen existiert somit eine Funktion

$$\eta : \mathbb{R}^1 \times \mathbb{R}^1 \times \mathbb{R}^k \to \mathbb{R}^{n-1} : (\xi, \lambda, \alpha) \mapsto \eta(\xi, \lambda, \alpha)$$

mit $\eta(0, \lambda^*, \alpha^*) = 0$ und

$$0 \equiv W_2\, F(x^* + V_2\eta(\xi, \lambda, \alpha) + v_1\xi\,,\, \lambda\,,\, \alpha). \tag{2.20}$$

in einer Umgebung von $(0, \lambda^*, \alpha^*)$.

Definition 2.13. Eine lokal definierte Funktion

$$g : \mathbb{R}^1 \times \mathbb{R}^1 \times \mathbb{R}^k \to \mathbb{R}^1 : (\xi, \lambda, \alpha) \mapsto g(\xi, \lambda, \alpha) := w_1^T\, F(x^* + V_2\eta(\xi, \lambda, \alpha) + v_1\xi, \lambda, \alpha)$$

heißt Ljapunov-Schmidt-reduzierte Funktion, wobei η implizit durch (2.20) definiert ist.

Es existieren folgende Beziehungen zwischen den von Jepson/Spence [JS84] verwendeten Vektoren bzw. Matrizen und den von Beyn [Bey84] verwendeten Projektoren:

$$W_2 \hat{=} M_{m,m-q}^T \begin{bmatrix} Q_r & Q_s \end{bmatrix}^{-1}, \quad w_1^T \hat{=} N_{m,q}^T \begin{bmatrix} Q_r & Q_s \end{bmatrix}^{-1}, \quad V_2 \hat{=} P_r\,, \quad v_1 \hat{=} P_s\,.$$

Griewank/Reddien [GR89] nutzten ein anderes System für eine Reihe singulärer Punkte. Beyn [Bey90] fand heraus, daß diese Beschreibung ebenfalls eine Ljapunov-Schmidt-Reduktion definiert. Im Fall $n = m$ oder $n = m + 1$ und $p = k = q = 1$ betrachteten Griewank/Reddien das System

$$F(x, \lambda, \alpha) = Dg\,, \quad Tx = \xi\,, \tag{2.21}$$

wobei $D \in \mathbb{R}^n$ und $T \in \mathbb{R}^{(n-m+1)\times n}$ so gewählt sind, daß

$$\begin{bmatrix} \partial_x F(x^*, \lambda^*, \alpha^*) & -D \\ T & 0 \end{bmatrix} \text{ regulär ist,} \tag{2.22}$$

was wegen Satz 2.4(a) und (b) im Fall $\mp Q_s = D$ und $\delta = 0$ erfüllt ist, falls $U_s^T D \neq 0$ und $T V_s$ regulär ist. Unter der Annahme von (2.22) liefert die Anwendung des Satzes über die impliziten Funktionen auf (2.21) lokale Funktionen

$$\begin{aligned} x &: \mathbb{R}^{n-m+1} \times \mathbb{R}^1 \times \mathbb{R}^1 \to \mathbb{R}^n : (\xi, \lambda, \alpha) \mapsto x(\xi, \lambda, \alpha) \qquad \text{und} \\ g &: \mathbb{R}^{n-m+1} \times \mathbb{R}^1 \times \mathbb{R}^1 \to \mathbb{R}^1 : (\xi, \lambda, \alpha) \mapsto g(\xi, \lambda, \alpha)\,, \end{aligned}$$

für die

$$\begin{array}{cc} F(x(\xi,\lambda,\alpha),\lambda,\alpha) \equiv Dg(\xi,\lambda,\alpha)\,, & Tx(\xi,\lambda,\alpha) \equiv \xi \quad \text{und} \\ x(Tx^*,\lambda^*,\alpha^*) = x^*, & g(Tx^*,\lambda^*,\alpha^*) = 0\,. \end{array} \tag{2.23}$$

gelten.

Bemerkung 2.14. Die durch (2.23) definierte Funktion g ist ebenfalls eine Ljapunov-Schmidt-reduzierte Funktion, wobei mit den Bezeichnungen von Beyn [Bey84]

$$D \hat{=} Q_s,\ \ T = N^T_{n,n-m+q} \begin{bmatrix} P_r & P_s \end{bmatrix}^{-1},\ \text{d. h.,}\ T\,P_r = 0 \in \mathbb{R}^{(n-m+q)\times(m-q)} \text{ und } T\,P_s = I_{n-m+q}$$

gilt, da

$$\begin{aligned} \begin{bmatrix} Q_r & 0 \end{bmatrix} \begin{bmatrix} Q_r & D \end{bmatrix}^{-1} F(x(\xi,\lambda,\alpha),\lambda,\alpha) &\equiv \begin{bmatrix} Q_r & 0 \end{bmatrix} \begin{bmatrix} Q_r & D \end{bmatrix}^{-1} Dg(\xi,\lambda,\alpha) \ \equiv\ 0 \\ N^T_{m,q} \begin{bmatrix} Q_r & D \end{bmatrix}^{-1} F(x(\xi,\lambda,\alpha),\lambda,\alpha) &\equiv N^T_{m,q} \begin{bmatrix} Q_r & D \end{bmatrix}^{-1} Dg(\xi,\lambda,\alpha) \ \equiv\ g(\xi,\lambda,\alpha) \\ x(Tx^*,\lambda^*,\alpha^*) &= x^* \\ N^T_{n,n-m+q} \begin{bmatrix} P_r & P_s \end{bmatrix}^{-1} x = Tx &= \xi\,. \end{aligned}$$

Govaerts [Gov97a] verallgemeinerte das System von Griewank/Reddien [GR89] zur Charakterisierung einer größeren Anzahl von singulären Punkten im Fall $n = m$ und $p = q = 1$. Damit Ableitungen der reduzierten Funktion in definierenden Systemen für singuläre Punkte verwendet werden können, definierte Govaerts die reduzierte Funktion für alle Punkte (x,λ,α) in einer Umgebung dieses singulären Punktes (x^*,λ^*,α^*). Unter der Voraussetzung (2.22) betrachtete Govaerts [Gov97a] das System

$$F(x+v,\lambda+\mu,\alpha+\beta) + Dg = F(x,\lambda,\alpha)\,, \quad Tv = \xi, \tag{2.24}$$

mit dem die impliziten Funktionen $v : \mathbb{R}^1 \times \mathbb{R}^1 \times \mathbb{R}^k \to \mathbb{R}^n : (\xi,\mu,\beta) \mapsto v(\xi,\mu,\beta)$ und $g : \mathbb{R}^1 \times \mathbb{R}^1 \times \mathbb{R}^k \to \mathbb{R}^1 : (\xi,\mu,\beta) \mapsto g(\xi,\mu,\beta)$ in einer Umgebung von $(0,0,0) \in \mathbb{R}^1 \times \mathbb{R}^1 \times \mathbb{R}^k$ definiert werden. Diese erfüllen

$$\begin{array}{cc} F(x+v(\xi,\mu,\beta),\lambda+\mu,\alpha+\beta) + Dg(\xi,\mu,\beta) \equiv F(x,\lambda,\alpha)\,, & Tv(\xi,\mu,\beta) \equiv \xi \quad \text{und} \\ v(0,0,0) = 0, & g(0,0,0) = 0 \end{array} \tag{2.25}$$

in dieser Umgebung.

Bemerkung 2.15.

(i) Griewank/Reddien [GR89] verwendeten in (2.21) $-Dg$, während Govaerts [Gov97a] im System (2.24) $+Dg$ nutzte. Dieser kleine Unterschied ist jedoch nicht bedeutsam.

(ii) Jeder Punkt (x,λ,α) in einer Umgebung von (x^*,λ^*,α^*) wird in den Koordinatenursprung bezüglich (ξ,μ,β) transformiert.

(iii) Das System (2.25) ermöglicht die Berechnung der Ableitungen von g für $(\xi,\mu,\beta) = (0,0,0)$ mittels impliziter Differentiation, wobei der Wert dieser Ableitungen nur von (x,λ,α) abhängt.

2.2.2 Zwei Definitionen der Ljapunov-Schmidt-Reduktion

In diesem Unterabschnitt sollen nun zwei allgemeine zueinander äquivalente Definitionen der Ljapunov-Schmidt-Reduktion eingeführt werden. Dabei werden Ideen aus dem vorhergehenden Unterabschnitt verwendet. Auf diese Art lassen sich auch Zusammenhänge zwischen den Vorgehensweisen erkennen, die im vorhergehenden Unterabschnitt vorgestellt wurden.

Eine Verallgemeinerung der Reduktionen von Beyn [Bey84] und von Jepson/Spence [JS84] ist die folgende Reduktion. Sie gilt für alle im Abschnitt 2.1 eingeführten Funktionen F und singulären Punkte $(x^*, \lambda^*, \alpha^*)$. Im Folgenden sei

$$U_s^T\, D \in \mathbb{R}^{q\times q} \quad \text{regulär} \tag{2.26}$$

und

$$T V_s \in \mathbb{R}^{(n-m+q)\times(n-m+q)} \quad \text{regulär,} \tag{2.27}$$

was im Fall $\pm Q_s = D$ der Aussage von Satz 2.4(b) entspricht. Es wird nun das System

$$0 = \begin{bmatrix} Q_r & 0 \end{bmatrix} \begin{bmatrix} Q_r & D \end{bmatrix}^{-1} F(x,\lambda,\alpha) \quad \text{mit} \quad \begin{bmatrix} \eta \\ \xi \end{bmatrix} = \begin{bmatrix} P_r & P_s \end{bmatrix}^{-1} (x - x^*) \tag{2.28}$$

mit den regulären Matrizen $\begin{bmatrix} Q_r & D \end{bmatrix}$ und $\begin{bmatrix} P_r & P_s \end{bmatrix}$ betrachtet, wobei $T \begin{bmatrix} P_r & P_s \end{bmatrix} = N^T_{n,n-m+q}$ ist. Im Fall $q = m$ bedeutet dies

$$0 = 0\,, \quad \xi = P_s^{-1}(x - x^*) \quad \text{und} \quad T = P_s^{-1}.$$

Lemma 2.16.
Unter der Voraussetzung, daß $q < m$ und $\begin{bmatrix} Q_r & D \end{bmatrix}$ regulär ist, ist die Forderung, daß

$$M^T_{m,m-q} \begin{bmatrix} Q_r & D \end{bmatrix}^{-1} \partial_x F(x^*,\lambda^*,\alpha^*)\, P_r \quad \textit{regulär ist,}$$

äquivalent zu Satz 2.4(f), wobei $W_2 = M^T_{m,m-q} \begin{bmatrix} Q_r & D \end{bmatrix}^{-1}$ gilt.

Bemerkung 2.17. $Q_r\, M^T_{m,m-q} \begin{bmatrix} Q_r & D \end{bmatrix}^{-1} = \begin{bmatrix} Q_r & 0 \end{bmatrix} \begin{bmatrix} Q_r & D \end{bmatrix}^{-1}$

Aus der Anwendung des Satzes über die impliziten Funktionen auf (2.28) in einer Umgebung von $(0, \lambda^*, \alpha^*)$ folgt im Fall $q < m$ die Existenz einer Funktion

$$\eta : \mathbb{R}^{n-m+q} \times \mathbb{R}^p \times \mathbb{R}^k \to \mathbb{R}^{m-q} : (\xi,\lambda,\alpha) \mapsto \eta(\xi,\lambda,\alpha)\,,$$

so daß

$$\begin{bmatrix} \eta(\xi,\lambda,\alpha) \\ \xi \end{bmatrix} = \begin{bmatrix} P_r & P_s \end{bmatrix}^{-1} (x(\xi,\lambda,\alpha) - x^*) \iff x(\xi,\lambda,\alpha) := x^* + P_r\,\eta(\xi,\lambda,\alpha) + P_s\,\xi\,, \tag{2.29}$$

wobei die Funktion x implizit durch

$$0 \equiv \begin{bmatrix} Q_r & 0 \end{bmatrix} \begin{bmatrix} Q_r & D \end{bmatrix}^{-1} F(x(\xi,\lambda,\alpha),\lambda,\alpha) \tag{2.30}$$

definiert ist. Dabei gilt

$$\eta(0,\lambda^*,\alpha^*) = 0\,, \quad \text{d.h.} \quad x(0,\lambda^*,\alpha^*) = x^*. \tag{2.31}$$

Im Fall $q = m$ geht dies in

$$x(\xi,\lambda,\alpha) := x^* + P_s\,\xi \quad \text{mit} \quad x(0,\lambda^*,\alpha^*) = x^* \tag{2.32}$$

über. Mit dieser implizit definierten Funktion $x(\xi,\lambda,\alpha)$ wird nun eine verallgemeinerte Ljapunov-Schmidt-reduzierte Funktion definiert.

Definition 2.18. Eine lokal definierte Funktion

$$g : \mathbb{R}^{n-m+q} \times \mathbb{R}^p \times \mathbb{R}^k \to \mathbb{R}^q : (\xi,\lambda,\alpha) \mapsto g(\xi,\lambda,\alpha) := \tilde{D}^T F(x(\xi,\lambda,\alpha),\lambda,\alpha) \tag{2.33}$$

ist eine verallgemeinerte Ljapunov-Schmidt-reduzierte Funktion, wobei die Matrix $\tilde{D} \in \mathbb{R}^{m\times q}$ gemäß

$$\tilde{D}^T D = \pm I_q \tag{2.34}$$

definiert ist.

Diese verallgemeinerte Ljapunov-Schmidt-reduzierte Funktion (2.33) kann auch durch die Verallgemeinerung der Idee von Griewank/Reddien [GR89] eingeführt werden. Dazu wird das System

$$F(x\,,\lambda,\alpha) \mp D\,g = c\,, \quad T(x - x^*) = \xi \tag{2.35}$$

betrachtet. Wegen (2.26) und (2.27) ergibt sich aus dem Satz 2.4(a) und (b) im Fall $Q_s = -D$ und $\delta = 0$, daß die Jacobimatrix

$$\begin{bmatrix} \partial_x F(x^*,\lambda^*,\alpha^*) & \mp D \\ T & 0 \end{bmatrix}$$

des Systems (2.35) bezüglich (x, g) im singulären Punkt regulär ist.

Bemerkung 2.19. Das Vorzeichen vor $D\,g$ sei so gewählt, daß

$$-D\,g \iff \tilde{D}^T D = +I_q\,, \quad \text{d.h.,} \quad +D\,g \iff \tilde{D}^T D = -I_q\,.$$

Aus dem Satz über die impliziten Funktionen folgt dann die Existenz lokaler Funktionen

$$\begin{aligned} \boldsymbol{x} &: \mathbb{R}^{n-m+q} \times \mathbb{R}^p \times \mathbb{R}^k \times \mathbb{R}^m \to \mathbb{R}^n : (\xi,\lambda,\alpha,c) \mapsto \boldsymbol{x}(\xi,\lambda,\alpha,c) \qquad \text{und} \\ \boldsymbol{g} &: \mathbb{R}^{n-m+q} \times \mathbb{R}^p \times \mathbb{R}^k \times \mathbb{R}^m \to \mathbb{R}^q : (\xi,\lambda,\alpha,c) \mapsto \boldsymbol{g}(\xi,\lambda,\alpha,c) \end{aligned}$$

für die in einer Umgebung von $(0,\lambda^*,\alpha^*,0)$

$$\begin{gathered} F(\boldsymbol{x}(\xi,\lambda,\alpha,c),\lambda,\alpha) \mp D\,\boldsymbol{g}(\xi,\lambda,\alpha,c) \equiv c\,, \quad T(\boldsymbol{x}(\xi,\lambda,\alpha,c) - x^*) \equiv \xi\,, \\ \boldsymbol{x}(0,\lambda^*,\alpha^*,0) = x^* \qquad \text{und} \qquad \boldsymbol{g}(0,\lambda^*,\alpha^*,0) = 0 \end{gathered} \tag{2.36}$$

gelten. Der Zusammenhang zu den in (2.29), (2.30) und (2.31) bzw. (2.32) und (2.33) definierten Funktionen wird in der folgenden Proposition beschrieben.

Proposition 2.20.
In einer Umgebung von $(0, \lambda^*, \alpha^*)$ *gilt*

$$\boldsymbol{x}(\xi, \lambda, \alpha, 0) \equiv x(\xi, \lambda, \alpha) \quad \textit{und} \quad \boldsymbol{g}(\xi, \lambda, \alpha, 0) \equiv g(\xi, \lambda, \alpha) \,.$$

Beweis. Die Funktionen x, g bzw. $\boldsymbol{x}$, $\boldsymbol{g}$ sind eindeutig durch (2.29), (2.30), (2.31) bzw. im Fall $q = m$ (2.32), (2.33) bzw. (2.36) definiert. Somit ist nur zu zeigen, daß $\boldsymbol{x}(\xi, \lambda, \alpha, 0)$ und $\boldsymbol{g}(\xi, \lambda, \alpha, 0)$ (2.29), (2.30), (2.31) und (2.33) erfüllt. Im Fall $q = m$ ist dann auch (2.32) erfüllt.

$$\begin{aligned}
N_{n,n-m+q}^T \begin{bmatrix} P_r & P_s \end{bmatrix}^{-1} (\boldsymbol{x}(\xi, \lambda, \alpha, 0) - x^*) &\equiv T(\boldsymbol{x}(\xi, \lambda, \alpha, 0) - x^*) \equiv \xi \\
\begin{bmatrix} Q_r & 0 \end{bmatrix} \begin{bmatrix} Q_r & D \end{bmatrix}^{-1} F(\boldsymbol{x}(\xi, \lambda, \alpha, 0), \lambda, \alpha) &\equiv \pm \begin{bmatrix} Q_r & 0 \end{bmatrix} \begin{bmatrix} Q_r & D \end{bmatrix}^{-1} D\,\boldsymbol{g}(\xi, \lambda, \alpha, 0) \equiv 0 \\
\boldsymbol{x}(0, \lambda^*, \alpha^*, 0) &= x^* \\
\tilde{D}^T F(\boldsymbol{x}(\xi, \lambda, \alpha, 0), \lambda, \alpha) &\equiv \pm \tilde{D}^T D\,\boldsymbol{g}(\xi, \lambda, \alpha, 0) \equiv \boldsymbol{g}(\xi, \lambda, \alpha, 0)
\end{aligned}$$

□

Bemerkung 2.21.

(i) Das Vorzeichen + oder − in (2.34) steht im unmittelbaren Zusammenhang mit dem gewählten Vorzeichen − oder + in (2.35), vergleiche die Bemerkung 2.19 und den obigen Beweis.

(ii) Die implizit definierten Funktionen $\boldsymbol{x}$ und $\boldsymbol{g}$ hängen von D und T, jedoch nicht von Q_r, der speziellen Wahl von $\begin{bmatrix} P_r & P_s \end{bmatrix}$ bei Beachtung von Bemerkung 2.14 und von der gemäß (2.34) gewählten Matrix $\tilde{D}$ ab.

(iii) Aus diesem Grund kann $Q_r := U_r$, $\tilde{D} := \pm U_s\,(D^T U_s)^{-1}$ und $P_s = V_s\,(T V_s)^{-1}$ ohne Beschränkung der Allgemeinheit gewählt werden. Dabei müssen U_r, U_s und V_s nicht bekannt sein.

(iv) Somit ist im Fall $n = m$ und $(x^*, \alpha^*) = (0, 0)$ die reduzierte Funktion von Beyn [Bey84] keine Verallgemeinerung der reduzierten Funktion von Golubitsky/Schaeffer [GS85]. Das bedeutet, daß für jede Funktion $\boldsymbol{g}(w, \alpha)$ gemäß Definition 2.10 eine Funktion $g(\xi, \alpha)$ gemäß Definition 2.8 mit $Q_s g(\xi, \alpha) = \boldsymbol{g}(P_s \xi, \alpha)$ existiert. Dabei tritt λ nicht auf.

(v) System (2.36) definiert eine verallgemeinerte Ljapunov-Schmidt-Reduktion. Die Ljapunov-Schmidt-Reduktion von Golubitsky/Schaeffer [GS85] ergibt sich im Spezialfall $n = m$, $p = 1$, $(x^*, \lambda^*, \alpha^*) = (0, 0, 0)$ und $c = 0$.

2.2.3 Verallgemeinerte Kontaktäquivalenz der reduzierten Funktionen

In diesem Abschnitt wird eine Äquivalenzrelation betrachtet, die eine Verallgemeinerung der Kontaktäquivalenz aus [GS85, Seite 166] ist. Deshalb wird sie verallgemeinerte Kontaktäquivalenz genannt. Anschließend werden reduzierte Funktionen eingeführt, die einer Verallgemeinerung von (2.36) genügen. Es wird gezeigt, daß diese reduzierten Funktionen äquivalent im Sinne der verallgemeinerten Kontaktäquivalenz sind.

Definition 2.22. Zwei glatte Funktionen

$$\boldsymbol{g} : (\boldsymbol{\xi}, \lambda, \alpha) \mapsto \boldsymbol{g}(\boldsymbol{\xi}, \lambda, \alpha) \in \mathbb{R}^q \quad \text{und} \quad g : (\xi, \lambda, \alpha) \mapsto g(\xi, \lambda, \alpha) \in \mathbb{R}^q$$

heißen äquivalent bezüglich der verallgemeinerten Kontaktäquivalenz, falls glatte Funktionen

$$S : (\boldsymbol{\xi}, \lambda, \alpha) \mapsto S(\boldsymbol{\xi}, \lambda, \alpha) \in \mathbb{R}^{q \times q} \quad \text{und} \quad \Xi : (\boldsymbol{\xi}, \lambda, \alpha) \mapsto \Xi(\boldsymbol{\xi}, \lambda, \alpha) \in \mathbb{R}^{n-m+q}$$

existieren, so daß $\Xi(0, \lambda^*, \alpha^*) = 0$, die Matrizen $S(0, \lambda^*, \alpha^*)$ und $\partial_{\boldsymbol{\xi}} \Xi(0, \lambda^*, \alpha^*)$ regulär sind und in einer Umgebung von $(0, \lambda^*, \alpha^*)$

$$\boldsymbol{g}(\boldsymbol{\xi}, \lambda, \alpha) = S(\boldsymbol{\xi}, \lambda, \alpha)\, g(\Xi(\boldsymbol{\xi}, \lambda, \alpha), \lambda, \alpha)$$

gilt.

Bemerkung 2.23.

(i) Folgende Beziehungen bestehen zu den Äquivalenzrelationen von Golubitsky/Schaeffer [GS85].

- In der Definition 1.2 in [GS85, Seite 398 im Kapitel IX] wird der Fall $n = m$, $p = 1$, $k = 0$, $\lambda^* = 0$ und $q > 1$ betrachtet. Dabei wird die Existenz einer Funktion Λ mit $\Lambda(\lambda^*) = \lambda^*$ und $\partial\Lambda(\lambda^*) > 0$ gefordert, so daß

  $$\boldsymbol{g}(\boldsymbol{\xi}, \lambda) = S(\boldsymbol{\xi}, \lambda)\, g(\Xi(\boldsymbol{\xi}, \lambda), \Lambda(\lambda))$$

  gilt.
- In der Definition auf Seite 5 in [GS85] wird zusätzlich $q = 1$, $S(0,0) > 0$ und $\partial_{\boldsymbol{\xi}} \Xi(0,0) > 0$ vorausgesetzt. Unter diesen Voraussetzungen werden dort Informationen über die Stabilität der Lösung hergeleitet.
- Wie in der Definition 2.22 wird in der Definition der starken Äquivalenz in [GS85, Seite 51] keine Funktion Λ verwendet. Die anderen Voraussetzungen entsprechen jedoch denen in der Definition auf Seite 5.
- Die Definition der Kontaktäquivalenz in [GS85, Seite 166] ist die Anwendung der Definition 2.22 auf den Spezialfall $n = m$ und $p = k = 0$.

(ii) In der Definition der Kontaktäquivalenz von Kunkel [Kun91] wird der Fall $n \geq m$, $k = 0$ und $\lambda^* = 0$ behandelt. Auch hier wird die Existenz einer Funktion Λ mit $\Lambda(0) = 0$ und mit der regulären Jacobimatrix $\partial\Lambda(0)$ gefordert. Ebenfalls wird bei dieser Kontaktäquivalenz verlangt, daß Funktionen Ξ und S mit $\Xi(0,0) = 0$ existieren, wobei die Matrizen $\partial_{\boldsymbol{\xi}} \Xi(0,0)$ und $S(0,0)$ regulär sind und

$$\boldsymbol{g}(\boldsymbol{\xi}, \lambda) = S(\boldsymbol{\xi}, \lambda)\, g(\Xi(\boldsymbol{\xi}, \lambda), \Lambda(\lambda))$$

gilt.

Lemma 2.24.
Die in Definition 2.22 definierte verallgemeinerten Kontaktäquivalenz ist eine Äquivalenzrelation.

Beweis. Es ist zu zeigen, daß diese Relation reflexiv, symmetrisch und transitiv ist. Die Transitivität ist offensichtlich erfüllt. Die Reflexivität $\boldsymbol{g}(\boldsymbol{\xi},\lambda,\alpha) = g(\boldsymbol{\xi},\lambda,\alpha)$ erhält man mit der Wahl $S(\boldsymbol{\xi},\lambda,\alpha) = I_q$ und $\Xi(\boldsymbol{\xi},\lambda,\alpha) = \boldsymbol{\xi}$. Mit der Wahl $\boldsymbol{S}(\xi,\lambda,\alpha) := S(\boldsymbol{\Xi}(\xi,\lambda,\alpha),\lambda,\alpha)^{-1}$, wobei $\boldsymbol{\Xi}$ die inverse Funktion von Ξ ist, ist auch die Symmetrie

$$\boldsymbol{g}(\boldsymbol{\xi},\lambda,\alpha) = S(\boldsymbol{\xi},\lambda,\alpha)\, g(\Xi(\boldsymbol{\xi},\lambda,\alpha),\lambda,\alpha) \Leftrightarrow g(\xi,\lambda,\alpha) = \boldsymbol{S}(\xi,\lambda,\alpha)\, \boldsymbol{g}(\boldsymbol{\Xi}(\xi,\lambda,\alpha),\lambda,\alpha)$$

gegeben. □

Jetzt wird gezeigt, daß Funktionen, die mit Hilfe einer Verallgemeinerung des Systems (2.36) definiert sind, zueinander äquivalent bezüglich der verallgemeinerten Kontaktäquivalenz, d. h. im Sinne von Definition 2.22, sind. Der Beweis erfolgt dabei analog zu bekannten Beweisen, vergleiche z. B. Golubitsky/Schaeffer [GS85, Satz 3.2 auf S. 31 und S. 47–50].

Proposition 2.25.
Die Funktion g aus

$$(2.37)\quad F(x(\xi,\lambda,\alpha,c),\lambda,\alpha) \mp D\, g(\xi,\lambda,\alpha,c) \equiv c\,,\ T(x(\xi,\lambda,\alpha,c) - x^*) \mp \delta\, g(\xi,\lambda,\alpha,c) \equiv \xi$$

mit $x(0,\lambda^*,\alpha^*,0) = x^*$ und $g(0,\lambda^*,\alpha^*,0) = 0$ und die Funktion $\boldsymbol{g}$ aus

$$(2.38)\quad F(\boldsymbol{x}(\boldsymbol{\xi},\lambda,\alpha,c),\lambda,\alpha) \mp D\, \boldsymbol{g}(\boldsymbol{\xi},\lambda,\alpha,c) \equiv c\,,\ \boldsymbol{T}(\boldsymbol{x}(\boldsymbol{\xi},\lambda,\alpha,c) - x^*) \mp \boldsymbol{\delta}\, \boldsymbol{g}(\boldsymbol{\xi},\lambda,\alpha,c) \equiv \boldsymbol{\xi}$$

mit $\boldsymbol{x}(0,\lambda^*,\alpha^*,0) = x^*$ und $\boldsymbol{g}(0,\lambda^*,\alpha^*,0) = 0$ sind äquivalent bezüglich der verallgemeinerten Kontaktäquivalenz.

Vor dem Beweis sollen einige Anmerkungen erfolgen.

Bemerkung 2.26.

(i) Die Regularitätsbedingungen (2.26) und (2.27) sind genau dann erfüllt, wenn $\begin{bmatrix} \partial_x F(x^*,\lambda^*,\alpha^*) & \mp D \\ T & \mp\delta \end{bmatrix}$ regulär ist, vergleiche Satz 2.4(a) im Fall $Q_s = D$. Mit dem Satz über die impliziten Funktionen folgt dann die lokale Existenz und Glattheit der Funktionen x und g.

(ii) Analog ist genau dann $\boldsymbol{T}\, V_s$ regulär und (2.26) erfüllt, wenn $\begin{bmatrix} \partial_x F(x^*,\lambda^*,\alpha^*) & \mp D \\ \boldsymbol{T} & \mp\boldsymbol{\delta} \end{bmatrix}$ regulär ist. In diesem Fall sind auch $\boldsymbol{x}$ und $\boldsymbol{g}$ lokal definiert und glatt.

(iii) Bei $\delta = 0 \in \mathbb{R}^{(n-m+q)\times q}$ ist System (2.37) mit System (2.36) identisch.

(iv) Die Vorzeichen + oder − werden gleichzeitig in (2.37) und (2.38) verwendet, d. h., es sind nur zwei unterschiedliche Fälle zu betrachten.

(v) Entsprechende Aussagen wie in dieser und der folgenden Proposition wie auch im folgenden Satz sind für Spezialfälle vom Fall $c = 0$ in der Literatur bekannt.

(vi) Um Definition 2.22 anwenden zu können, wird in der Proposition 2.25, in ihrem Beweis und in weiteren analogen Anwendungen c als Teil des des Parametervektors α betrachtet.

Beweis. Die Mengen $(x(\xi,\lambda,\alpha,c),\lambda,\alpha,g(\xi,\lambda,\alpha,c))$ und $(\boldsymbol{x}(\boldsymbol{\xi},\lambda,\alpha,c),\lambda,\alpha,\boldsymbol{g}(\boldsymbol{\xi},\lambda,\alpha,c))$ sind Lösungsmengen $(x,\lambda,\alpha,g)\in\mathbb{R}^{n+p+k+q}$ von

$$F(x,\lambda,\alpha)\mp D\,g\equiv c \tag{2.39}$$

mit der Dimension $n-m+q+p+k+m=n+p+k+q$. Umgekehrt lassen sich auch (c,ξ) aus (2.37) bzw. $(c,\boldsymbol{\xi})$ aus (2.38) als Funktionen von (x,λ,α,g) bzw. $(\boldsymbol{x},\lambda,\alpha,\boldsymbol{g})$ darstellen. Deshalb sind die Funktionen $(x(\xi,\lambda,\alpha,c),\lambda,\alpha,g(\xi,\lambda,\alpha,c))$ von (ξ,λ,α,c) und $(\boldsymbol{x}(\boldsymbol{\xi},\lambda,\alpha,c),\lambda,\alpha,\boldsymbol{g}(\boldsymbol{\xi},\lambda,\alpha,c))$ von $(\boldsymbol{\xi},\lambda,\alpha,c)$ invertierbar. Somit enthalten beide oben definierten Mengen alle Lösungen (x,λ,α,g) von (2.39), d. h., beide Mengen sind gleich. Damit existiert eine Funktion Ξ mit

$$(x(\Xi(\boldsymbol{\xi},\lambda,\alpha,c),\lambda,\alpha,c),\lambda,\alpha,g(\Xi(\boldsymbol{\xi},\lambda,\alpha,c),\lambda,\alpha,c))=(\boldsymbol{x}(\boldsymbol{\xi},\lambda,\alpha,c),\lambda,\alpha,\boldsymbol{g}(\boldsymbol{\xi},\lambda,\alpha,c))\,,$$

woraus

$$\begin{aligned}\Xi(\boldsymbol{\xi},\lambda,\alpha,c)&\equiv T(x(\Xi(\boldsymbol{\xi},\lambda,\alpha,c),\lambda,\alpha,c)-x^*)\mp\delta\,g(\Xi(\boldsymbol{\xi},\lambda,\alpha,c),\lambda,\alpha,c)\\&\equiv T(\boldsymbol{x}(\boldsymbol{\xi},\lambda,\alpha,c)-x^*)\mp\delta\,\boldsymbol{g}(\boldsymbol{\xi},\lambda,\alpha,c)\\&\equiv(T-\boldsymbol{T})(\boldsymbol{x}(\boldsymbol{\xi},\lambda,\alpha,c)-x^*)\mp(\delta-\boldsymbol{\delta})\,\boldsymbol{g}(\boldsymbol{\xi},\lambda,\alpha,c)+\boldsymbol{\xi}\,.\end{aligned}$$

folgt. Im singulären Punkt ergibt sich damit

$$\begin{aligned}\Xi(0,\lambda^*,\alpha^*,0)&=T(\boldsymbol{x}(0,\lambda^*,\alpha^*,0)-x^*)\mp\delta\,\boldsymbol{g}(0,\lambda^*,\alpha^*,0)\\&=T(x^*-x^*)\mp\delta\,0\\&=0\\\partial_{\boldsymbol{\xi}}\Xi(0,\lambda^*,\alpha^*,0)&=T\partial_{\boldsymbol{\xi}}\boldsymbol{x}(0,\lambda^*,\alpha^*,0)\mp\delta\,\partial_{\boldsymbol{\xi}}\boldsymbol{g}(0,\lambda^*,\alpha^*,0)\\&=\begin{bmatrix}T & \mp\delta\end{bmatrix}\begin{bmatrix}\partial_x F(x^*,\lambda^*,\alpha^*) & \mp D\\ \boldsymbol{T} & \mp\boldsymbol{\delta}\end{bmatrix}^{-1}N_{n+q,n-m+q}\,.\end{aligned}$$

Da

$$\begin{bmatrix}\partial_x F(x^*,\lambda^*,\alpha^*) & \mp D\\ T & \mp\delta\end{bmatrix}\begin{bmatrix}\partial_x F(x^*,\lambda^*,\alpha^*) & \mp D\\ \boldsymbol{T} & \mp\boldsymbol{\delta}\end{bmatrix}^{-1}N_{n+q,n-m+q}$$

spaltenregulär und

$$\begin{bmatrix}\partial_x F(x^*,\lambda^*,\alpha^*) & \mp D\end{bmatrix}\begin{bmatrix}\partial_x F(x^*,\lambda^*,\alpha^*) & \mp D\\ \boldsymbol{T} & \mp\boldsymbol{\delta}\end{bmatrix}^{-1}N_{n+q,n-m+q}=0$$

ist, ist die Matrix

$$\partial_{\boldsymbol{\xi}}\Xi(0,\lambda^*,\alpha^*,0)\in\mathbb{R}^{(n-m+q)\times(n-m+q)}$$

regulär. Wegen der Gleichheit $\boldsymbol{g}(\boldsymbol{\xi},\lambda,\alpha,c)=g(\Xi(\boldsymbol{\xi},\lambda,\alpha,c),\lambda,\alpha,c)$ und der Glattheit von Ξ sind die Funktionen $\boldsymbol{g}$ und g somit äquivalent bezüglich der verallgemeinerten Kontaktäquivalenz. □

Bemerkung 2.27.

(i) Im Fall $T=\boldsymbol{T}$ gilt speziell $\Xi(\boldsymbol{\xi},\lambda,\alpha,c)=\mp(\delta-\boldsymbol{\delta})\,\boldsymbol{g}(\boldsymbol{\xi},\lambda,\alpha,c)+\boldsymbol{\xi}$.

(ii) Im Fall $\delta=\boldsymbol{\delta}$ erhält man dagegen $\Xi(\boldsymbol{\xi},\lambda,\alpha,c)=(T-\boldsymbol{T})(\boldsymbol{x}(\boldsymbol{\xi},\lambda,\alpha,c)-x^*)+\boldsymbol{\xi}$.

Proposition 2.28.
Die Funktion g aus

$$(2.40)\quad F(x(\boldsymbol{\xi},\lambda,\alpha,c),\lambda,\alpha) \mp D\,g(\boldsymbol{\xi},\lambda,\alpha,c) \equiv c,\ \boldsymbol{T}(x(\boldsymbol{\xi},\lambda,\alpha,c)-x^*) \mp \boldsymbol{\delta}\,g(\boldsymbol{\xi},\lambda,\alpha,c) \equiv \boldsymbol{\xi}$$

mit $x(0,\lambda^*,\alpha^*,0)=x^*$ *und* $g(0,\lambda^*,\alpha^*,0)=0$ *und die Funktion* $\boldsymbol{g}$ *aus*

$$(2.41)\quad F(\boldsymbol{x}(\boldsymbol{\xi},\lambda,\alpha,c),\lambda,\alpha) \mp \boldsymbol{D}\,\boldsymbol{g}(\boldsymbol{\xi},\lambda,\alpha,c) \equiv c,\ \boldsymbol{T}(\boldsymbol{x}(\boldsymbol{\xi},\lambda,\alpha,c)-x^*) \mp \boldsymbol{\delta}\,\boldsymbol{g}(\boldsymbol{\xi},\lambda,\alpha,c) \equiv \boldsymbol{\xi}$$

mit $\boldsymbol{x}(0,\lambda^*,\alpha^*,0)=x^*$ *und* $\boldsymbol{g}(0,\lambda^*,\alpha^*,0)=0$ *sind äquivalent bezüglich der verallgemeinerten Kontaktäquivalenz.*

Obwohl die Bezeichnungen der impliziten Funktionen $(\boldsymbol{x},\boldsymbol{g})$ in (2.38) sich von den Bezeichnungen der impliziten Funktionen (x,g) in (2.40) unterscheiden, sind beide Funktionenpaare identisch. Diese unterschiedlichen Bezeichnungen ergeben sich daraus, daß die impliziten Funktionen g aus (2.37) und $\boldsymbol{g}$ aus (2.41) im Satz 2.29 verwendet werden.

Beweis. Wegen (2.26) und falls $\boldsymbol{T}V_s$ und $U_s^T\boldsymbol{D}$ regulär ist, werden die Systeme

$$(2.42)\qquad \begin{aligned} F(x(\boldsymbol{\xi},\lambda,\alpha,c,g),\lambda,\alpha) \mp D\,g + D\,\nu(\boldsymbol{\xi},\lambda,\alpha,c,g) &\equiv c\,,\\ \boldsymbol{T}(x(\boldsymbol{\xi},\lambda,\alpha,c,g)-x^*) \mp \boldsymbol{\delta}\,g + \boldsymbol{\delta}\,\nu(\boldsymbol{\xi},\lambda,\alpha,c,g) &\equiv \boldsymbol{\xi}\,,\\ x(0,\lambda^*,\alpha^*,0,0) &= x^*,\\ \nu(0,\lambda^*,\alpha^*,0,0) &= 0 \end{aligned}$$

und

$$(2.43)\qquad \begin{aligned} F(\boldsymbol{x}(\boldsymbol{\xi},\lambda,\alpha,c,\nu),\lambda,\alpha) \mp \boldsymbol{D}\,\boldsymbol{g}(\boldsymbol{\xi},\lambda,\alpha,c,\nu) + D\,\nu &\equiv c\,,\\ \boldsymbol{T}(\boldsymbol{x}(\boldsymbol{\xi},\lambda,\alpha,c,\nu)-x^*) \mp \boldsymbol{\delta}\,\boldsymbol{g}(\boldsymbol{\xi},\lambda,\alpha,c,\nu) + \boldsymbol{\delta}\,\nu &\equiv \boldsymbol{\xi}\,,\\ \boldsymbol{x}(0,\lambda^*,\alpha^*,0,0) &= x^*,\\ \boldsymbol{g}(0,\lambda^*,\alpha^*,0,0) &= 0\,. \end{aligned}$$

betrachtet. Wenn die Funktion $\nu(\boldsymbol{\xi},\lambda,\alpha,c,g)$ aus (2.42) in (2.43) eingesetzt wird, ergibt sich

(2.44)

$$\begin{aligned} F(\boldsymbol{x}(\boldsymbol{\xi},\lambda,\alpha,c,\nu(\boldsymbol{\xi},\lambda,\alpha,c,g)),\lambda,\alpha) \mp \boldsymbol{D}\,\boldsymbol{g}(\boldsymbol{\xi},\lambda,\alpha,c,\nu(\boldsymbol{\xi},\lambda,\alpha,c,g)) + D\,\nu(\boldsymbol{\xi},\lambda,\alpha,c,g) &\equiv c,\\ \boldsymbol{T}(\boldsymbol{x}(\boldsymbol{\xi},\lambda,\alpha,c,\nu(\boldsymbol{\xi},\lambda,\alpha,c,g))-x^*) \mp \boldsymbol{\delta}\,\boldsymbol{g}(\boldsymbol{\xi},\lambda,\alpha,c,\nu(\boldsymbol{\xi},\lambda,\alpha,c,g)) + \boldsymbol{\delta}\,\nu(\boldsymbol{\xi},\lambda,\alpha,c,g) &\equiv \boldsymbol{\xi},\\ \boldsymbol{x}(0,\lambda^*,\alpha^*,0,\nu(0,\lambda^*,\alpha^*,0,0)) &= x^*\\ \text{und}\qquad \boldsymbol{g}(0,\lambda^*,\alpha^*,0,\nu(0,\lambda^*,\alpha^*,0,0)) &= 0. \end{aligned}$$

Ein Vergleich der Systeme (2.42) mit (2.44) jeweils mit $g=0$ liefert

$$\boldsymbol{g}(\boldsymbol{\xi},\lambda,\alpha,c,\nu(\boldsymbol{\xi},\lambda,\alpha,c,0)) \equiv 0$$

und somit

$$\begin{aligned} \boldsymbol{g}(\boldsymbol{\xi},\lambda,\alpha,c,\nu(\boldsymbol{\xi},\lambda,\alpha,c,g)) &= \boldsymbol{g}(\boldsymbol{\xi},\lambda,\alpha,c,\nu(\boldsymbol{\xi},\lambda,\alpha,c,g)) - \boldsymbol{g}(\boldsymbol{\xi},\lambda,\alpha,c,\nu(\boldsymbol{\xi},\lambda,\alpha,c,0))\\ &= \int_0^1 d_{tg}\boldsymbol{g}(\boldsymbol{\xi},\lambda,\alpha,c,\nu(\boldsymbol{\xi},\lambda,\alpha,c,tg))\,g\,dt\\ &= \int_0^1 d_{tg}\boldsymbol{g}(\boldsymbol{\xi},\lambda,\alpha,c,\nu(\boldsymbol{\xi},\lambda,\alpha,c,tg))\,dt\,g\\ &=: \boldsymbol{S}(\boldsymbol{\xi},\lambda,\alpha,c,\nu(\boldsymbol{\xi},\lambda,\alpha,c,g))\,g\,. \end{aligned}$$

Offensichtlich ist $\boldsymbol{S}$ glatt. Weiterhin ist (2.40) identisch mit

$$
\begin{aligned}
F(x(\boldsymbol{\xi},\lambda,\alpha,c),\lambda,\alpha) \mp D\,(g(\boldsymbol{\xi},\lambda,\alpha,c) \pm \nu) + D\,\nu &\equiv c,\\
\boldsymbol{T}(x(\boldsymbol{\xi},\lambda,\alpha,c) - x^*) \mp \boldsymbol{\delta}\,(g(\boldsymbol{\xi},\lambda,\alpha,c) \pm \nu) + \boldsymbol{\delta}\,\nu &\equiv \boldsymbol{\xi},\\
x(0,\lambda^*,\alpha^*,0) &= x^*,\\
g(0,\lambda^*,\alpha^*,0) &= 0.
\end{aligned}
\tag{2.45}
$$

Durch Vergleich mit (2.42) ergibt sich die Beziehung

$$
g = g(\boldsymbol{\xi},\lambda,\alpha,c) \pm \nu \quad\Longleftrightarrow\quad \nu(\boldsymbol{\xi},\lambda,\alpha,c,g) = \nu
$$

und damit

$$
\boldsymbol{g}(\boldsymbol{\xi},\lambda,\alpha,c,\nu) = \boldsymbol{S}(\boldsymbol{\xi},\lambda,\alpha,c,\nu)\,(g(\boldsymbol{\xi},\lambda,\alpha,c) \pm \nu)\,.
$$

Daraus folgt

$$
\partial_\nu \boldsymbol{g}(\boldsymbol{\xi},\lambda,\alpha,c,\nu) = \partial_\nu \boldsymbol{S}(\boldsymbol{\xi},\lambda,\alpha,c,\nu)\,(g(\boldsymbol{\xi},\lambda,\alpha,c) \pm \nu) + \boldsymbol{S}(\boldsymbol{\xi},\lambda,\alpha,c,\nu)\,\partial_\nu(\pm\nu)
$$

und speziell

$$
\begin{aligned}
\partial_\nu \boldsymbol{g}(0,\lambda^*,\alpha^*,0,0) &= \partial_\nu \boldsymbol{S}(0,\lambda^*,\alpha^*,0,0)\,g(0,\lambda^*,\alpha^*,0) \pm \boldsymbol{S}(0,\lambda^*,\alpha^*,0,0)\\
&= \pm \boldsymbol{S}(0,\lambda^*,\alpha^*,0,0)
\end{aligned}
$$

mit

$$
\begin{aligned}
\partial_\nu \boldsymbol{g}(0,\lambda^*,\alpha^*,0,0) &= N^T_{n+q,q} \begin{bmatrix} \partial_\nu \boldsymbol{x}(0,\lambda^*,\alpha^*,0,0) \\ \partial_\nu \boldsymbol{g}(0,\lambda^*,\alpha^*,0,0) \end{bmatrix}\\
&= N^T_{n+q,q} \begin{bmatrix} \partial_x F(x^*,\lambda^*,\alpha^*) & \mp \boldsymbol{D} \\ \boldsymbol{T} & \mp \boldsymbol{\delta} \end{bmatrix}^{-1} \begin{bmatrix} -D \\ -\boldsymbol{\delta} \end{bmatrix}.
\end{aligned}
$$

Da

$$
N^T_{n+q,q} \begin{bmatrix} \partial_x F(x^*,\lambda^*,\alpha^*) & \mp \boldsymbol{D} \\ \boldsymbol{T} & \mp \boldsymbol{\delta} \end{bmatrix}^{-1} \begin{bmatrix} \partial_x F(x^*,\lambda^*,\alpha^*) & -D \\ \boldsymbol{T} & -\boldsymbol{\delta} \end{bmatrix}
$$

zeilenregulär und

$$
N^T_{n+q,q} \begin{bmatrix} \partial_x F(x^*,\lambda^*,\alpha^*) & \mp \boldsymbol{D} \\ \boldsymbol{T} & \mp \boldsymbol{\delta} \end{bmatrix}^{-1} \begin{bmatrix} \partial_x F(x^*,\lambda^*,\alpha^*) \\ \boldsymbol{T} \end{bmatrix} = 0
$$

ist, ist die Matrix

$$
\pm \boldsymbol{S}(0,\lambda^*,\alpha^*,0,0) = \partial_\nu \boldsymbol{g}(0,\lambda^*,\alpha^*,0,0) \in \mathbb{R}^{q\times q}
$$

regulär. Somit ist $\boldsymbol{g}$ aus (2.43) und $(g \pm \nu)$ aus (2.45) äquivalent bezüglich der verallgemeinerten Kontaktäquivalenz. Dies gilt auch für $\nu = 0$. In diesem Fall wird (2.43) zu (2.41) und (2.45) zu (2.40), woraus die verallgemeinerte Kontaktäquivalenz von g aus (2.40) und $\boldsymbol{g}$ aus (2.41) folgt. □

Wenn die Proposition 2.25 mit der Proposition 2.28 zusammengefaßt wird, ergibt sich

Satz 2.29.
Die Funktion g aus (2.37) und die Funktion $\boldsymbol{g}$ aus (2.41) sind äquivalent bezüglich der verallgemeinerten Kontaktäquivalenz.

2.2.4 Einführung einer Ableitungsäquivalenz-Relation

Wie in einer Reihe anderer Arbeiten, z. B. [PS81] und [GR84], sollen auch in dieser Arbeit singuläre Punkte mittels erweiterter Systeme unter Anwendung des Newtonverfahrens bestimmt werden. Zur Verringerung des Aufwands soll dabei das minimal erweiterte System (2.10) verwendet werden, das im Spezialfall $l = 0$ das minimal erweiterte System (2.8) ist. Zur effizienten Lösung des linearisierten Systems ist es sinnvoll, statt der reduzierten Funktion g aus (2.37) eine reduzierte Funktion zu verwenden, in deren definierendem System die Matrix $\bar{D} \in \mathbb{R}^{m\times l}$ statt der Matrix $D \in \mathbb{R}^{m\times q}$ verwendet wird. So kann die Dimension des geränderten Systems (2.37) auf den kleinstmöglichen Wert reduziert werden. Weiterhin ist so erreichbar, daß die ersten m Zeilen der Jacobimatrix des erweiterten Systems und des geränderten Systems übereinstimmen. Dann tritt in allen zu lösenden linearen Systemen dieselbe Koeffizientenmatrix, die Jacobimatrix des geränderten Systems, auf.

Aus diesem Grund wird im Folgenden eine andere reduzierte Funktion, eine Parameterfunktion, mit den gewünschten Eigenschaften beschrieben. Von dieser Funktion wird vorausgesetzt, daß sie im gewissen Sinne äquivalent zur Funktion g aus (2.37) ist, d. h., diese Funktion soll die gleichen definierenden Gleichungen und Nichtentartungsbedingungen wie g erfüllen.

Bevor diese Parameterfunktion eingeführt wird, erfolgt die Definition dieser Äquivalenzrelation. Dazu werden zuerst folgende Relationen definiert.

Definition 2.30. Für zwei Funktionen

$$(2.46)\qquad \begin{array}{ll} g : \mathbb{R}^{n-m+q}\times\mathbb{R}^{r} \times \mathbb{R}^{p+k+l-q} \times \mathbb{R}^{m} \to \mathbb{R}^{q} : (\xi,\vartheta,\tau,c) \mapsto g(\xi,\vartheta,\tau,c) & \text{und} \\ h : \mathbb{R}^{n-m+q}\times \qquad \mathbb{R}^{p+k+l-q} \times \mathbb{R}^{m} \to \mathbb{R}^{q} : (\xi, \quad \tau,c) \mapsto h(\xi,\tau,c) & \end{array} \quad \text{gilt:}$$

$g\,(R1)\,h \quad \Leftrightarrow$ es existiert ein ausgezeichneter Punkt (ξ,ϑ,τ,c), so daß gilt

$$\partial_{(\xi,\tau)} g(\xi,\vartheta,\tau,c)a = 0 \Longleftrightarrow \partial_{(\xi,\tau)} h(\xi,\tau,c)a = 0 \quad \text{für alle} \quad a \in \mathbb{R}^{n+p+k+l-m}.$$

$g\,(R2)\,h \quad \Leftrightarrow$ es existiert ein ausgezeichneter Punkt (ξ,ϑ,τ,c), so daß für alle Mengen $\{a_1, a_2, \ldots, a_i\}$ von Vektoren $a_\iota \in \mathbb{R}^{n+p+k+l-m}$

$$\partial^i_{(\xi,\tau)} g(\xi,\vartheta,\tau,c)\,[a_1,\ldots,a_i] = 0 \quad \Longleftrightarrow \quad \partial^i_{(\xi,\tau)} h(\xi,\tau,c)\,[a_1,\ldots,a_i] = 0$$

gilt, falls $\partial^j_{(\xi,\tau)} g(\xi,\vartheta,\tau,c)[a_{i_1}, a_{i_2}, \ldots, a_{i_j}] = 0$ für alle Untermengen $\{a_{i_1}, a_{i_2}, \ldots, a_{i_j}\} \subset \{a_1, a_2, \ldots, a_i\}$, $1 \le j < i, 1 \le i_1 < i_2 < \ldots < i_j \le i$ erfüllt ist.

$g\,(R3)\,h \quad \Leftrightarrow$ im Fall $q = 1$, $n + p + k + l \ge m + 2$ existiert ein ausgezeichneter Punkt (ξ,ϑ,τ,c), so daß für alle Paare von Komponenten (ζ_1,ζ_2) von (ξ,τ) mit $\partial_{\zeta_1} g(\xi,\vartheta,\tau,c) = 0$ und $D_i(g)_{(\zeta_1,\zeta_2)}(\xi,\vartheta,\tau,c) = 0$ für alle $1 \le i \le j$

$$D_{j+1}(g)_{(\zeta_1,\zeta_2)}(\xi,\vartheta,\tau,c) = 0 \quad \Longleftrightarrow \quad D_{j+1}(h)_{(\zeta_1,\zeta_2)}(\xi,\tau,c) = 0$$

gilt, wobei $D_1(g)_{(\zeta_1,\zeta_2)} := \partial_{\zeta_2} g$ und für alle $j \ge 1$:

$$D_{j+1}(g)_{(\zeta_1,\zeta_2)} := (\partial^2_{\zeta_1} g)\partial_{\zeta_2} D_j(g)_{(\zeta_1,\zeta_2)} - (\partial_{\zeta_1}\partial_{\zeta_2} g)(\partial_{\zeta_1} D_j(g)_{(\zeta_1,\zeta_2)})\,.$$

$g\,(R4)\,h \quad \Leftrightarrow$ im Fall $q = 1$, $n + p + k + l \geq m + 2$ existiert ein ausgezeichneter Punkt $(\xi, \vartheta, \tau, c)$, so daß für alle Paare von Komponenten (ζ_1, ζ_2) von (ξ, τ) mit $\partial_\zeta g(\xi, \vartheta, \tau, c) = 0$ und $\partial^2_\zeta g(\xi, \vartheta, \tau, c) = 0$, wobei $\zeta := (\zeta_1, \zeta_2)^T$ ist,

$$\operatorname{sgn} D_\infty(g)_\zeta(\xi, \vartheta, \tau, c) = \operatorname{sgn} D_\infty(h)_\zeta(\xi, \tau, c)$$

gilt, wobei

$$\begin{aligned} D_\infty(g)_\zeta := {} & \left(\partial^3_{\zeta_1} g \partial^3_{\zeta_2} g - \partial^2_{\zeta_1} \partial_{\zeta_2} g \partial_{\zeta_1} \partial^2_{\zeta_2} g\right)^2 - \\ & - 4\left(\partial^3_{\zeta_1} g \partial_{\zeta_1} \partial^2_{\zeta_2} g - (\partial^2_{\zeta_1} \partial_{\zeta_2} g)^2\right)\left(\partial^2_{\zeta_1} \partial_{\zeta_2} g \partial^3_{\zeta_2} g - (\partial_{\zeta_1} \partial^2_{\zeta_2} g)^2\right) \end{aligned}$$

die Diskriminante der homogenen quadratischen Gleichung $\det\left(\partial^3_\zeta g \begin{bmatrix} \varrho \\ \sigma \end{bmatrix}\right) = 0$ bezüglich ϱ und σ ist, vergleiche Govaerts[Gov95].

$g\,(R5)\,h \quad \Leftrightarrow$ im Fall $q = 2$, $n = m$, $p + k + l \geq 2$ existiert ein ausgezeichneter Punkt $(\xi, \vartheta, \tau, c)$, so daß, falls $\partial_\xi g(\xi, \vartheta, \tau, c) = 0$, dann auch

$$\operatorname{sgn} D_{2,2}(g)(\xi, \vartheta, \tau, c) = \operatorname{sgn} D_{2,2}(h)(\xi, \tau, c)$$

gilt, wobei $D_{2,2}(g)$ die Diskriminante der homogenen quadratischen Gleichung $\det\left(\partial^2_\xi g \begin{bmatrix} \varrho \\ \sigma \end{bmatrix}\right) = 0$ bezüglich ϱ und σ ist, vergleiche Govaerts[Gov95] und $g\,(R4)\,h$.

$g\,(R6)\,h \quad \Leftrightarrow$ im Fall $q = 2$, $n = m$, $k + l \geq 2$, $p = 1$ und λ ist keine Komponente von ϑ existiert ein ausgezeichneter Punkt $(\xi, \vartheta, \tau, c)$, so daß, falls $\partial_\xi g(\xi, \vartheta, \tau, c) = 0$ und $D_{2,2}(g)(\xi, \vartheta, \tau, c) > 0$, auch

$$\partial_\lambda g(\xi, \vartheta, \tau, c) = \partial^2_\xi g(\xi, \vartheta, \tau, c) \begin{bmatrix} \varrho_1 \\ \sigma_1 \end{bmatrix} \begin{bmatrix} \varrho_1 \\ \sigma_1 \end{bmatrix} t_1 + \partial^2_\xi g(\xi, \vartheta, \tau, c) \begin{bmatrix} \varrho_2 \\ \sigma_2 \end{bmatrix} \begin{bmatrix} \varrho_2 \\ \sigma_2 \end{bmatrix} t_2$$

und

$$\partial_\lambda h(\xi, \tau, c) = \partial^2_\xi h(\xi, \tau, c) \begin{bmatrix} \varrho_1 \\ \sigma_1 \end{bmatrix} \begin{bmatrix} \varrho_1 \\ \sigma_1 \end{bmatrix} t_1 + \partial^2_\xi h(\xi, \tau, c) \begin{bmatrix} \varrho_2 \\ \sigma_2 \end{bmatrix} \begin{bmatrix} \varrho_2 \\ \sigma_2 \end{bmatrix} t_2$$

für die Lösungsmenge $\{t \begin{bmatrix} \varrho_i \\ \sigma_i \end{bmatrix} : \varrho_i, \sigma_i \text{ fest}, \ t \in \mathbb{R}, i = 1, 2\}$ der Gleichung

$$\det\left(\partial^2_\xi g(\xi, \vartheta, \tau, c) \begin{bmatrix} \varrho \\ \sigma \end{bmatrix}\right) = 0 \tag{2.47}$$

und Zahlen t_1 und t_2 gilt.

Vor der Definition der Ableitungsäquivalenz sollen einige Bemerkungen zur Relation $g\,(R6)\,h$ erfolgen.

Bemerkung 2.31.

(i) Wegen $D_{2,2}(g)(\xi, \vartheta, \tau, c) > 0$ hat die Gleichung (2.47) zwei sich schneidende Geraden $\{t \begin{bmatrix} \varrho_i \\ \sigma_i \end{bmatrix} : \varrho_i, \sigma_i \text{ fest}, t \in \mathbb{R}, i = 1, 2\}$ als Lösungsmenge.

(ii) Nach [GS85, S. 403] gilt $\partial^2_\xi g(\xi, \vartheta, \tau, c) \begin{bmatrix} \varrho_i \\ \sigma_i \end{bmatrix} \begin{bmatrix} \varrho_i \\ \sigma_i \end{bmatrix} \neq 0 \in \mathbb{R}^2$ für $i = 1, 2$.

(iii) Es gilt bei Unterdrückung der Argumente $(\xi, \vartheta, \tau, c)$:

$$
\begin{aligned}
t_1 = t_2 = 0 &\Leftrightarrow \partial_\lambda g = 0 \in \mathbb{R}^2 \\
t_1 \neq 0, t_2 = 0 &\Leftrightarrow \partial_\lambda g \parallel \partial_\xi^2 g \begin{bmatrix} \varrho_1 \\ \sigma_1 \end{bmatrix} \begin{bmatrix} \varrho_1 \\ \sigma_1 \end{bmatrix} \\
t_1 = 0, t_2 \neq 0 &\Leftrightarrow \partial_\lambda g \parallel \partial_\xi^2 g \begin{bmatrix} \varrho_2 \\ \sigma_2 \end{bmatrix} \begin{bmatrix} \varrho_2 \\ \sigma_2 \end{bmatrix} \\
t_1 > 0, t_2 > 0 &\Rightarrow \partial_\lambda g \in \operatorname{cone} \left\{ \partial_\xi^2 g \begin{bmatrix} \varrho_1 \\ \sigma_1 \end{bmatrix} \begin{bmatrix} \varrho_1 \\ \sigma_1 \end{bmatrix}, \partial_\xi^2 g \begin{bmatrix} \varrho_2 \\ \sigma_2 \end{bmatrix} \begin{bmatrix} \varrho_2 \\ \sigma_2 \end{bmatrix} \right\} \\
t_1 > 0, t_2 < 0 &\Rightarrow \partial_\lambda g \in \operatorname{cone} \left\{ \partial_\xi^2 g \begin{bmatrix} \varrho_1 \\ \sigma_1 \end{bmatrix} \begin{bmatrix} \varrho_1 \\ \sigma_1 \end{bmatrix}, -\partial_\xi^2 g \begin{bmatrix} \varrho_2 \\ \sigma_2 \end{bmatrix} \begin{bmatrix} \varrho_2 \\ \sigma_2 \end{bmatrix} \right\} \\
t_1 < 0, t_2 < 0 &\Rightarrow \partial_\lambda g \in \operatorname{cone} \left\{ -\partial_\xi^2 g \begin{bmatrix} \varrho_1 \\ \sigma_1 \end{bmatrix} \begin{bmatrix} \varrho_1 \\ \sigma_1 \end{bmatrix}, -\partial_\xi^2 g \begin{bmatrix} \varrho_2 \\ \sigma_2 \end{bmatrix} \begin{bmatrix} \varrho_2 \\ \sigma_2 \end{bmatrix} \right\} \\
t_1 < 0, t_2 > 0 &\Rightarrow \partial_\lambda g \in \operatorname{cone} \left\{ -\partial_\xi^2 g \begin{bmatrix} \varrho_1 \\ \sigma_1 \end{bmatrix} \begin{bmatrix} \varrho_1 \\ \sigma_1 \end{bmatrix}, \partial_\xi^2 g \begin{bmatrix} \varrho_2 \\ \sigma_2 \end{bmatrix} \begin{bmatrix} \varrho_2 \\ \sigma_2 \end{bmatrix} \right\}
\end{aligned}
$$

Die letzten vier Relationen werden ebenfalls zur Äquivalenzrelation, falls jeweils das Innere der Kegel, d. h. die offenen Kegel, betrachtet werden.

Definition 2.32. Die Funktionen g und h sind äquivalent bezüglich der Ableitungsäquivalenz, d. h., g ist äquivalent zu h und h ist äquivalent zu g, falls

- im Fall $q = 1$, $n+p+k+l \geq m+2$ für die Funktionen g und h die Relationen $g\,(R1)\,h$, $g\,(R2)\,h$, $g\,(R3)\,h$ und $g\,(R4)\,h$ für den selben Punkt $(\xi, \vartheta, \tau, c)$ gelten.
- im Fall, daß $m = n$, $q = 2$, $k + l \geq 2$, $p = 1$ und λ keine Komponente von ϑ ist, für die Funktionen g und h die Relationen $g\,(R1)\,h$, $g\,(R2)\,h$, $g\,(R5)\,h$ und $g\,(R6)\,h$ für den selben Punkt $(\xi, \vartheta, \tau, c)$ gelten.
- im Fall, daß $m = n$, $q = 2$, $p + k + l \geq 2$ und entweder $p = 1$ und λ eine Komponente von ϑ oder $p \neq 1$ ist, für die Funktionen g und h die Relationen $g\,(R1)\,h$, $g\,(R2)\,h$ und $g\,(R5)\,h$ für den selben Punkt $(\xi, \vartheta, \tau, c)$ gelten.
- in allen übrigen Fällen für die Funktionen g und h die Relationen $g\,(R1)\,h$ und $g\,(R2)\,h$ für den selben Punkt $(\xi, \vartheta, \tau, c)$ gelten.

Die Definition 2.32 wird durch folgende Bemerkungen motiviert.

Bemerkung 2.33.

(i) Die Ableitungsäquivalenz wurde so gewählt, daß die definierenden Gleichungen und Nichtentartungsbedingungen für die einfachsten singulären Punkte bei äquivalenten Funktionen im Sinne der Definition 2.32, d. h. bezüglich der Ableitungsäquivalenz, gleichzeitig erfüllt bzw. gleichzeitig nicht erfüllt sind. Diese definierenden Gleichungen und Nichtentartungsbedingungen wurden von Govaerts [Gov95], [Gov97a], [Gov97b] aufgelistet. Die definierenden Gleichungen und Nichtentartungsbedingungen haben gerade die Eigenschaft, daß bezüglich der verallgemeinerten Kontaktäquivalenz äquivalente Funktionen diese gleichzeitig erfüllen bzw. nicht erfüllen. Deshalb ist in gewissem Sinne Definition 2.32 eine Verallgemeinerung der Definition 2.22.

(ii) Somit lassen sich analog zu den Klassifikationen von Golubitsky/Schaeffer [GS85], Jepson/Spence [JS84] bzw. Govaerts [Gov95] und [Gov97a] auch mit der ableitungs-

äquivalenten Funktion h die einfachsten singulären Punkte klassifizieren. Diese Klassifikation erfolgt mittels der im Unterabschnitt 2.2.5 definierten zu g ableitungsäquivalenten Funktion $\boldsymbol{\vartheta}$ im Abschnitt 2.3.

(iii) Gemischte Ableitungen in $g\,(R2)\,h$ werden dadurch charakterisiert, daß die a_i Koordinatenvektoren sind, die aber zu unterschiedlichen Koordinaten gehören, während bei reinen Ableitungen alle a_i die gleichen Koordinatenvektoren sind.

(iv) Für uns ist nur der Fall $r = q$ interessant. Wenn aber nicht auch $r = 0$ zugelassen wird, ist die Relation nicht reflexiv und damit keine Äquivalenzrelation.

(v) In den in dieser Arbeit betrachteten Fällen ist der ausgezeichnete Punkt $(\xi, \vartheta, \tau, c)$ der singuläre Punkt $(0, \vartheta^*, \tau^*, 0)$. Dann gilt $\partial_\xi g(\xi, \vartheta, \tau, c) = 0$, woraus für ableitungsäquivalente Funktionen h wegen $g\,(R1)\,h$ sofort $\partial_\xi h(\xi, \tau, c) = 0$ folgt.

(vi) Falls für ein $j \in \{1, \ldots, n+p+k+l-m\}$ ein i_j existiert, so daß

$$\partial^i_{(\xi,\tau)} g(\xi, \vartheta, \tau, c) \left[e^j\right]^i = 0 \quad \forall i = 1, \ldots, i_j - 1 \quad \text{und} \quad \partial^{i_j}_{(\xi,\tau)} g(\xi, \vartheta, \tau, c) \left[e^j\right]^{i_j} \neq 0$$

gilt, wobei $e^j \in \mathbb{R}^{n+p+k+l-m}$ der j-te Koordinatenvektor ist, und g und h ableitungsäquivalent sind, dann gilt für das selbe j und das selbe i_j

$$\partial^i_{(\xi,\tau)} h(\xi, \tau, c) \left[e^j\right]^i = 0 \quad \text{und} \quad \partial^{i_j}_{(\xi,\tau)} h(\xi, \tau, c) \left[e^j\right]^{i_j} \neq 0$$

für alle $i = 1, \ldots, i_j - 1$.

(vii) Für $j \in \{1, \ldots, n-m+q\}$ ist im betrachteten singulären Punkt somit $i_j > 1$.

(viii) Offensichtlich ist $D_2(g)_{(\zeta_1,\zeta_2)} = (\partial^2_{\zeta_1} g)(\partial^2_{\zeta_2} g) - (\partial_{\zeta_1}\partial_{\zeta_2} g)^2 = \det \partial^2_{(\zeta_1,\zeta_2)} g\,.$

Lemma 2.34.
Die Relationen in Definition 2.32 sind Äquivalenzrelationen.

Beweis. Mit einer geeigneten Wahl der Dimensionen sind die Relationen reflexiv, symmetrisch und transitiv. □

2.2.5 Die Anwendung der Ableitungsäquivalenz-Relation auf einen neuen Typ reduzierter Funktionen

In diesem Unterabschnitt soll nun die zu g aus (2.37) ableitungsäquivalente Parameterfunktion definiert und deren Äquivalenz gezeigt werden.

Dazu wird die Funktion

$$F(x, \lambda, \alpha) + \bar{D}\,\mu$$

mit $\bar{D}$ aus (2.9) betrachtet. Im Fall $l = 0$ ist dies gerade die Funktion (2.2). Weiterhin steht diese Funktion in allen Fällen $l \geq 0$ in den ersten m Zeilen der linken Seite des erweiterten Systems (2.8) bzw. (2.10). Offensichtlich gilt für diese Funktion

$$F(x^*, \lambda^*, \alpha^*) + \bar{D} \cdot 0 = 0, \quad \begin{aligned} &\operatorname{rank} \partial_x \left(F(x^*, \lambda^*, \alpha^*) + \bar{D} \cdot 0\right) &&= m - q \quad \text{und} \\ &\operatorname{rank} \partial_{(x,\lambda,\alpha,\mu)} \left(F(x^*, \lambda^*, \alpha^*) + \bar{D} \cdot 0\right) &&= m. \end{aligned}$$

Durch Umordnung und Neubezeichnung der Parameter (λ, α, μ) erhält man die Parameter $\vartheta \in \mathbb{R}^q$ und $\tau \in \mathbb{R}^{p+k+l-q}$, so daß

$$(2.48) \qquad \begin{array}{ll} \boldsymbol{F}(x,\vartheta,\tau) & := F(x,\lambda,\alpha) + \bar{D}\,\mu, \\ \boldsymbol{F}(x^*,\vartheta^*,\tau^*) & = 0 \end{array} \quad \text{und} \quad \begin{array}{ll} \operatorname{rank} \partial_x \boldsymbol{F}(x^*,\vartheta^*,\tau^*) & = m-q, \\ \operatorname{rank} \partial_{(x,\vartheta)} \boldsymbol{F}(x^*,\vartheta^*,\tau^*) & = m, \end{array}$$

wobei (ϑ^*, τ^*) gerade $(\lambda^*, \alpha^*, 0)$ entspricht.

Wegen Proposition 2.28, (2.26), Satz 2.4 und (2.9) kann o. B. d. A. nun D so gewählt werden, daß $\operatorname{im} \bar{D} \subseteq \operatorname{im} D$ ist. Im Fall $q = l$ ist somit $\operatorname{im} \bar{D} = \operatorname{im} D$. Eine verallgemeinerte Ljapunov-Schmidt-reduzierte Funktion gemäß Definition 2.18 der Funktion $\boldsymbol{F}$ genügt dann mit Umordnung der Argumente

$$\boldsymbol{g}(\xi,\vartheta,\tau) = g(\xi,\lambda,\alpha) + \tilde{D}^T \bar{D}\,\mu\,,$$

wobei g gemäß (2.33) definiert ist. Im Fall $l = 0$ stimmt somit $\boldsymbol{g}$ mit g überein. Wegen Proposition 2.20 läßt sich die Funktion $\boldsymbol{g}$ auch mittels

$$(2.49) \qquad \begin{array}{c} \boldsymbol{F}(\boldsymbol{x}(\xi,\vartheta,\tau),\vartheta,\tau) \mp D\,\boldsymbol{g}(\xi,\vartheta,\tau) \equiv 0\,, \quad T(\boldsymbol{x}(\xi,\vartheta,\tau) - x^*) \equiv \xi\,, \\ \boldsymbol{x}(0,\vartheta^*,\tau^*) = x^* \qquad \text{und} \qquad \boldsymbol{g}(0,\vartheta^*,\tau^*) = 0 \end{array}$$

definieren, was mit der Umordnung der Parameter gerade

$$\begin{array}{c} F(x(\xi,\lambda,\alpha),\lambda,\alpha) + \bar{D}\,\mu \mp D\left(g(\xi,\lambda,\alpha) + \tilde{D}^T \bar{D}\,\mu\right) \equiv 0\,, \quad T(x(\xi,\lambda,\alpha) - x^*) \equiv \xi\,, \\ x(0,\lambda^*,\alpha^*) = x^* \qquad \text{und} \qquad g(0,\lambda^*,\alpha^*) = 0 \end{array}$$

bedeutet. Bei Einführung des Parametervektors c wie in (2.36) kann man die Funktionen g und x gemäß

$$(2.50) \qquad \begin{array}{c} \boldsymbol{F}(x(\xi,\vartheta,\tau,c),\vartheta,\tau) \mp D\,g(\xi,\vartheta,\tau,c) \equiv c\,, \quad T(x(\xi,\vartheta,\tau,c) - x^*) \equiv \xi\,, \\ x(0,\vartheta^*,\tau^*,0) = x^* \qquad \text{und} \qquad g(0,\vartheta^*,\tau^*,0) = 0 \end{array}$$

definieren. Offensichtlich gilt für die Funktionen $\boldsymbol{x}$ und $\boldsymbol{g}$ aus (2.49) und x und g aus (2.50) in einer Umgebung von $(0, \vartheta^*, \tau^*)$

$$\boldsymbol{x}(\xi,\vartheta,\tau) \equiv x(\xi,\vartheta,\tau,0) \quad \text{und} \quad \boldsymbol{g}(\xi,\vartheta,\tau) \equiv g(\xi,\vartheta,\tau,0)\,.$$

Weiterhin ist in einer Umgebung von $(0, \vartheta^*, \tau^*, 0)$ wegen (2.50)

$$\begin{bmatrix} \partial_x \boldsymbol{F}(x(\xi,\vartheta,\tau,c),\vartheta,\tau) & \mp D \\ T & 0 \end{bmatrix} \begin{bmatrix} \partial_\vartheta x(\xi,\vartheta,\tau,c) \\ \partial_\vartheta g(\xi,\vartheta,\tau,c) \end{bmatrix} \equiv \begin{bmatrix} -\partial_\vartheta \boldsymbol{F}(x(\xi,\vartheta,\tau,c),\vartheta,\tau) \\ 0 \end{bmatrix},$$

wobei die Systemmatrix in dieser Umgebung regulär ist. Wegen (2.48) und (2.27) ist für $(\xi, \vartheta, \tau, c) = (0, \vartheta^*, \tau^*, 0)$ auch $\begin{bmatrix} \partial_x \boldsymbol{F}(x^*,\vartheta^*,\tau^*) & -\partial_\vartheta \boldsymbol{F}(x^*,\vartheta^*,\tau^*) \\ T & 0 \end{bmatrix}$ regulär, weshalb $N_{n+q,q}^T \begin{bmatrix} \partial_x \boldsymbol{F}(x^*,\vartheta^*,\tau^*) & \mp D \\ T & 0 \end{bmatrix}^{-1} \begin{bmatrix} \partial_x \boldsymbol{F}(x^*,\vartheta^*,\tau^*) & -\partial_\vartheta \boldsymbol{F}(x^*,\vartheta^*,\tau^*) \\ T & 0 \end{bmatrix}$ zeilenregulär ist. Da $N_{n+q,q}^T \begin{bmatrix} \partial_x \boldsymbol{F}(x^*,\vartheta^*,\tau^*) & \mp D \\ T & 0 \end{bmatrix}^{-1} \begin{bmatrix} \partial_x \boldsymbol{F}(x^*,\vartheta^*,\tau^*) \\ T \end{bmatrix} = 0 \in \mathbb{R}^{q \times n}$ ist, ist

$$(2.51) \quad N_{n+q,q}^T \begin{bmatrix} \partial_x \boldsymbol{F}(x^*,\vartheta^*,\tau^*) & \mp D \\ T & 0 \end{bmatrix}^{-1} \begin{bmatrix} \partial_x \boldsymbol{F}(x^*,\vartheta^*,\tau^*) & -\partial_\vartheta \boldsymbol{F}(x^*,\vartheta^*,\tau^*) \\ T & 0 \end{bmatrix} N_{n+q,q}$$
$$= \partial_\vartheta g(0,\vartheta^*,\tau^*,0) \qquad \text{regulär.}$$

Nach dem Satz über die impliziten Funktionen existiert damit eine lokale Funktion $\vartheta : \mathbb{R}^{n-m+q} \times \mathbb{R}^{p+k+l-q} \times \mathbb{R}^m \to \mathbb{R}^q : (\xi, \tau, c) \mapsto \vartheta(\xi, \tau, c)$ mit

$$\vartheta(0, \tau^*, 0) = \vartheta^* \quad \text{und} \quad g(\xi, \vartheta(\xi, \tau, c), \tau, c) = 0 \tag{2.52}$$

für alle (ξ, τ, c) in einer Umgebung von $(0, \tau^*, 0)$.

Satz 2.35.
Die implizit definierten Funktionen g aus (2.50) und ϑ aus (2.52) sind ableitungsäquivalent.

Der Beweis dieses Satzes erfolgt im Unterabschnitt 2.2.6.

Die Funktion ϑ läßt sich auch direkt mit der Funktion $\boldsymbol{F}$ bestimmen. Wenn die Funktion ϑ aus (2.52) in (2.50) eingesetzt wird, erhält man

$$\begin{aligned} \boldsymbol{F}(x(\xi, \vartheta(\xi, \tau, c), \tau, c), \vartheta(\xi, \tau, c), \tau) \mp D\, g(\xi, \vartheta(\xi, \tau, c), \tau, c) &\equiv c\,, \\ T(x(\xi, \vartheta(\xi, \tau, c), \tau, c) - x^*) &\equiv \xi\,, \\ g(\xi, \vartheta(\xi, \tau, c), \tau, c) &\equiv 0\,, \\ x(0, \vartheta^*, \tau^*, 0) = x^* \qquad &\text{und} \\ \vartheta(0, \tau^*, 0) = \vartheta^* \end{aligned}$$

und deshalb mit $x(\xi, \tau, c) := x(\xi, \vartheta(\xi, \tau, c), \tau, c)$

$$\begin{aligned} \boldsymbol{F}(x(\xi, \tau, c), \vartheta(\xi, \tau, c), \tau) \equiv c\,, \quad T(x(\xi, \tau, c) - x^*) \equiv \xi\,, \\ x(0, \tau^*, 0) = x^* \qquad \text{und} \qquad \vartheta(0, \tau^*, 0) = \vartheta^*, \end{aligned} \tag{2.53}$$

wodurch ϑ eindeutig bestimmt ist.

Bemerkung 2.36.

(i) Diese Funktion ϑ wurde von Schnabel [Sch94] zur Charakterisierung von einfachen, zweifachen und dreifachen Rückkehrpunkten verwendet. Dabei ist $m = n$, $p = q = 1$, $k = l = 0$ und somit $\vartheta = \lambda$, während τ nicht existiert. Zweifache Rückkehrpunkte werden auch als Hysteresepunkte bezeichnet. Bei der Charakterisierung von ein- und zweifachen Rückkehrpunkten stützte sich Schnabel dabei auf [PS81] und [Pön87].

(ii) In diesen Arbeiten wie auch in [PSS99] und [Pön90] wird jedoch statt

$$T(x(\xi, \tau, c) - x^*) \equiv \xi \quad \text{die Identität} \quad \boldsymbol{T} \begin{bmatrix} \boldsymbol{x}(\boldsymbol{\xi}, \tau, c) - x^* \\ \boldsymbol{\vartheta}(\boldsymbol{\xi}, \tau, c) - \vartheta^* \end{bmatrix} \equiv \boldsymbol{\xi}$$

verwendet, wobei $\boldsymbol{T} \begin{bmatrix} V_s \\ 0 \end{bmatrix}$ regulär ist. In diesen vier Arbeiten ist ebenfalls $m = n$, $p = q = 1$ und $l = 0$. In den ersten drei Arbeiten ist $\vartheta = \lambda$, während bei den in [Pön90] betrachteten (einfachen) Verzweigungspunkten $k = 1$, $\vartheta = \alpha$ und $\tau = \lambda$ gilt.

- Pönisch und Schwetlick [PS81] (1981) untersuchten den einfachen Rückkehrpunkt mit $k = 0$, weshalb τ nicht existiert.
- Pönisch [Pön87] (1987) betrachtete den Hysteresepunkt mit $k = 1$ und $\tau = \alpha$.
- Die Charakterisierungsmethode von Schnabel [Sch94] wurde von Pönisch, Schnabel und Schwetlick [PSS99] (1999) für beliebige Rückkehrpunkte der Vielfachheit $\ell > 1$ verallgemeinert. Im Fall $k = 0$ tritt auch dort der Parameter τ nicht auf, während im Fall $k > 0$ ebenfalls $\tau = \alpha$ gesetzt wurde.

Damit die Funktion $\boldsymbol{\vartheta}$ verwendet werden kann, ist noch zu zeigen, daß g aus (2.50) und $\boldsymbol{\vartheta}$ ableitungsäquivalent bzw. ϑ aus (2.52) bzw. (2.53) und $\boldsymbol{\vartheta}$ äquivalent bezüglich der verallgemeinerten Kontaktäquivalenz sind. Wegen (2.27) und der Proposition 2.25 kann o. B. d. A. deshalb

$$\boldsymbol{T} = \begin{bmatrix} T & T_2 \end{bmatrix}, \quad \text{d. h.,} \quad \boldsymbol{T}\, M_{n+q,n} = T \tag{2.54}$$

gewählt werden.

Satz 2.37.
Die Funktionen $\boldsymbol{\vartheta}$ aus

$$\begin{aligned} \boldsymbol{F}(\boldsymbol{x}(\boldsymbol{\xi},\tau,c),\boldsymbol{\vartheta}(\boldsymbol{\xi},\tau,c),\tau) \equiv c, \qquad \boldsymbol{T} \begin{bmatrix} \boldsymbol{x}(\boldsymbol{\xi},\tau,c) - x^* \\ \boldsymbol{\vartheta}(\boldsymbol{\xi},\tau,c) - \vartheta^* \end{bmatrix} \equiv \boldsymbol{\xi} \quad \textit{mit} \\ \boldsymbol{x}(0,\tau^*,0) = x^*, \qquad\qquad \boldsymbol{\vartheta}(0,\tau^*,0) = \vartheta^* \end{aligned} \tag{2.55}$$

und ϑ aus (2.53) bzw. (2.52) sind äquivalent bezüglich der verallgemeinerten Kontaktäquivalenz.

Beweis. Dieser Beweis erfolgt analog zum Beweis von der Proposition 2.25.
$(x(\xi,\tau,c),\vartheta(\xi,\tau,c),\tau)$ und $(\boldsymbol{x}(\boldsymbol{\xi},\tau,c),\boldsymbol{\vartheta}(\boldsymbol{\xi},\tau,c),\tau)$ sind die Mengen der Dimension $n-m+q+p+k+l-q+m = n+p+k+l$ aller Punkte, die

$$F(x,\vartheta,\tau) \equiv c$$

erfüllen. Somit sind sie gleich, d. h.,

$$\begin{aligned} (x(\Xi(\boldsymbol{\xi},\tau,c),\tau,c),\vartheta(\Xi(\boldsymbol{\xi},\tau,c),\tau,c),\tau) &= (\boldsymbol{x}(\boldsymbol{\xi},\tau,c),\boldsymbol{\vartheta}(\boldsymbol{\xi},\tau,c),\tau) \\ \Rightarrow \Xi(\boldsymbol{\xi},\tau,c) &\equiv T(x(\Xi(\boldsymbol{\xi},\tau,c),\tau,c) - x^*) \\ &\equiv T(\boldsymbol{x}(\boldsymbol{\xi},\tau,c) - x^*) \\ &\equiv \boldsymbol{\xi} - T_2(\boldsymbol{\vartheta}(\boldsymbol{\xi},\tau,c) - \vartheta^*) \\ \Xi(0,\tau^*,0) &= T(\boldsymbol{x}(0,\tau^*,0) - x^*) \qquad = 0 \\ \partial_{\boldsymbol{\xi}}\Xi(0,\tau^*,0) &= T\partial_{\boldsymbol{\xi}}\boldsymbol{x}(0,\tau^*,0) \\ &= \begin{bmatrix} T & 0 \end{bmatrix} \begin{bmatrix} \partial_{(x,\vartheta)}F(x^*,\vartheta^*,\tau^*) \\ \boldsymbol{T} \end{bmatrix}^{-1} N_{n+q,n-m+q}\,. \end{aligned}$$

Dabei ist

$$\partial_{\boldsymbol{\xi}}\Xi(0,\tau^*,0) \in \mathbb{R}^{(n-m+q)\times(n-m+q)}$$

regulär, da

$$\begin{bmatrix} \partial_x F(x^*,\vartheta^*,\tau^*) & \partial_\vartheta F(x^*,\vartheta^*,\tau^*) \\ T & 0 \end{bmatrix} \begin{bmatrix} \partial_{(x,\vartheta)}F(x^*,\vartheta^*,\tau^*) \\ \boldsymbol{T} \end{bmatrix}^{-1} N_{n+q,n-m+q}$$

spaltenregulär und

$$\partial_{(x,\vartheta)}F(x^*,\vartheta^*,\tau^*) \begin{bmatrix} \partial_{(x,\vartheta)}F(x^*,\vartheta^*,\tau^*) \\ \boldsymbol{T} \end{bmatrix}^{-1} N_{n+q,n-m+q} = 0$$

ist. Die Funktion Ξ ist glatt und $\boldsymbol{\vartheta}(\boldsymbol{\xi},\tau,c) = \vartheta(\Xi(\boldsymbol{\xi},\tau,c),\tau,c)$. Somit sind die Funktionen $\boldsymbol{\vartheta}$ und ϑ äquivalent bezüglich der verallgemeinerten Kontaktäquivalenz. □

2.2.6 Der Beweis von Satz 2.35

Beweis. Es werden die Funktionen g aus (2.50) und ϑ aus (2.52) betrachtet. Zu zeigen ist die Ableitungsäquivalenz dieser Funktionen. Dabei entspricht h der Funktion ϑ, in (2.46) ist $r = q$ und der ausgezeichnete Punkt ist der singuläre Punkt $(\xi, \vartheta, \tau, c) = (0, \vartheta^*, \tau^*, 0)$.

$g\,(R1)\,\vartheta$**:** Die Ableitung der zweiten Gleichung von (2.52) nach (ξ, τ) liefert

$$\partial_{(\xi,\tau)} g(\xi, \vartheta(\xi,\tau,c), \tau, c) \equiv -\partial_\vartheta g(\xi, \vartheta(\xi,\tau,c), \tau, c)\, \partial_{(\xi,\tau)} \vartheta(\xi,\tau,c)\,, \tag{2.56}$$

was im singulären Punkt

$$\partial_{(\xi,\tau)} g(0, \vartheta^*, \tau^*, 0) = -\partial_\vartheta g(0, \vartheta^*, \tau^*, 0)\, \partial_{(\xi,\tau)} \vartheta(0, \tau^*, 0)$$

bedeutet. Wegen (2.51) folgt für alle $a \in \mathbb{R}^{n+p+k+l-m}$

$$\partial_{(\xi,\tau)} g(0, \vartheta^*, \tau^*, 0)\,[a] = 0 \quad \Longleftrightarrow \quad \partial_{(\xi,\tau)} \vartheta(0, \tau^*, 0)\,[a] = 0\,.$$

$g\,(R2)\,\vartheta$**:** Gezeigt wird die Äquivalenz für $i = 2, 3, \ldots$ per Induktion.

Für $i = 2$ gilt für $\iota = 1, 2$

$$\partial_{(\xi,\tau)} g(0, \vartheta^*, \tau^*, 0)\,[a_\iota] = 0 \quad \text{und somit auch} \quad \partial_{(\xi,\tau)} \vartheta(0, \tau^*, 0)\,[a_\iota] = 0\,.$$

Die zweite Ableitung der zweiten Gleichung in (2.52) ergibt dann

$$\partial^2_{(\xi,\tau)} g(0, \vartheta^*, \tau^*, 0)\,[a_1, a_2] = -\partial_\vartheta g(0, \vartheta^*, \tau^*, 0)\, \partial^2_{(\xi,\tau)} \vartheta(0, \tau^*, 0)\,[a_1, a_2]\,,$$

woraus wegen (2.51)

$$\partial^2_{(\xi,\tau)} g(0, \vartheta^*, \tau^*, 0)\,[a_1, a_2] = 0 \quad \Longleftrightarrow \quad \partial^2_{(\xi,\tau)} \vartheta(0, \tau^*, 0)\,[a_1, a_2] = 0$$

folgt.

Für $i > 2$ werden die Terme

$$\partial^i_{(\xi,\tau)} g(0, \vartheta^*, \tau^*, 0)\,[a_1, \ldots, a_i] \quad \text{und} \quad \partial^i_{(\xi,\tau)} \vartheta(0, \tau^*, 0)\,[a_1, \ldots, a_i]$$

betrachtet. Da für alle j mit $1 \le j < i$ und alle Untermengen

$$\{a_{i_1}, a_{i_2}, \ldots, a_{i_j}\} \subset \{a_1, a_2, \ldots, a_i\},\ 1 \le j < i, 1 \le i_1 < i_2 < \ldots < i_j \le i$$

$$\partial^j_{(\xi,\tau)} g(0, \vartheta^*, \tau^*, 0)\,[a_{i_1}, \ldots, a_{i_j}] = 0$$

und somit auch

$$\partial^j_{(\xi,\tau)} \vartheta(0, \tau^*, 0)\,[a_{i_1}, \ldots, a_{i_j}] = 0$$

erfüllt sind, gilt für die i-te Ableitung der zweiten Gleichung in (2.52) im singulären Punkt

$$\partial^i_{(\xi,\tau)} g(0, \vartheta^*, \tau^*, 0)\,[a_1, \ldots, a_i] = -\partial_\vartheta g(0, \vartheta^*, \tau^*, 0)\, \partial^i_{(\xi,\tau)} \vartheta(0, \tau^*, 0)\,[a_1, \ldots, a_i]\,.$$

Dabei wurden alle verschwindenden Terme weggelassen. Wegen (2.51) ergibt sich

$$\partial^i_{(\xi,\tau)} g(0, \vartheta^*, \tau^*, 0)\,[a_1, \ldots, a_i] = 0 \quad \Longleftrightarrow \quad \partial^i_{(\xi,\tau)} \vartheta(0, \tau^*, 0)\,[a_1, \ldots, a_i] = 0\,,$$

woraus $g\,(R2)\,\vartheta$ folgt.

$g\,(R3)\,\vartheta$: Dieser Beweis wird analog zu den Beweisen von Govaerts [Gov97a] und [Gov95] geführt. Dort erfolgten die Beweise für einige Spezialfälle der in Definition 2.22 eingeführten verallgemeinerten Kontaktäquivalenz.

Es seien nun $q = 1$, $n+p+k+l \geq m+2$ und damit $g, \vartheta \in \mathbb{R}^1$ und $(\xi, \tau) \in \mathbb{R}^{n+p+k+l-m}$. Deshalb enthält (ξ, τ) mindestens zwei Komponenten. Zwei dieser Komponenten seien (ζ_1, ζ_2). Für diese beiden Komponenten wird $g\,(R3)\,\vartheta$ gezeigt. Zuerst wird dabei per Induktion

$$
\begin{aligned}
(2.57)\qquad & \begin{bmatrix} \partial_{\zeta_1} g \\ \partial_{\zeta_2} g = D_1(g)_{(\zeta_1,\zeta_2)} \\ \vdots \\ D_{j-1}(g)_{(\zeta_1,\zeta_2)} \\ D_j(g)_{(\zeta_1,\zeta_2)} \end{bmatrix} (\xi, \vartheta(\xi,\tau,c), \tau, c) \\
& = \begin{bmatrix} S_{0,0} & 0 & \dots & 0 & 0 \\ S_{1,0} & S_{1,1} & \ddots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ S_{j-1,0} & S_{j-1,1} & \dots & S_{j-1,j-1} & 0 \\ S_{j,0} & S_{j,1} & \dots & S_{j,j-1} & S_{j,j} \end{bmatrix} \begin{bmatrix} \partial_{\zeta_1} \vartheta \\ \partial_{\zeta_2} \vartheta = D_1(\vartheta)_{(\zeta_1,\zeta_2)} \\ \vdots \\ D_{j-1}(\vartheta)_{(\zeta_1,\zeta_2)} \\ D_j(\vartheta)_{(\zeta_1,\zeta_2)} \end{bmatrix} (\xi, \tau, c)
\end{aligned}
$$

nachgewiesen, wobei $S_{i,j}$ Kombinationen von Funktionen sind, die von $(\xi, \vartheta(\xi,\tau,c), \tau, c)$ bzw. (ξ, τ, c) abhängen. Dabei gilt für die Diagonalelemente der Matrix

$$
\begin{aligned}
(2.58)\qquad & S_{0,0}(\xi, \vartheta(\xi,\tau,c), \tau, c) = -\partial_\vartheta g(\xi, \vartheta(\xi,\tau,c), \tau, c) \qquad \text{und} \\
& S_{i,i}(\xi, \vartheta(\xi,\tau,c), \tau, c) = \left(-\partial_\vartheta g(\xi, \vartheta(\xi,\tau,c), \tau, c)\right)^i \quad \text{mit} \quad i = 1, 2, \dots, j.
\end{aligned}
$$

Für den singulären Punkt $(\xi, \tau, c) = (0, \tau^*, 0)$ kann daraus $g\,(R3)\,\vartheta$ gefolgert werden.

Aus (2.56) ergibt sich

$$
\begin{aligned}
& \begin{bmatrix} \partial_{\zeta_1} g \\ \partial_{\zeta_2} g = D_1(g)_{(\zeta_1,\zeta_2)} \end{bmatrix} (\xi, \vartheta(\xi,\tau,c), \tau, c) \\
& \qquad = \begin{bmatrix} -\partial_\vartheta g(\xi, \vartheta(\xi,\tau,c), \tau, c) & 0 \\ 0 & -\partial_\vartheta g(\xi, \vartheta(\xi,\tau,c), \tau, c) \end{bmatrix} \begin{bmatrix} \partial_{\zeta_1} \vartheta \\ \partial_{\zeta_2} \vartheta \end{bmatrix} (\xi, \tau, c),
\end{aligned}
$$

d. h., (2.57) ist für $j = 1$ gezeigt. Weiterhin ist

$$
\begin{aligned}
& S_{0,0}(\xi, \vartheta(\xi,\tau,c), \tau, c) = -\partial_\vartheta g(\xi, \vartheta(\xi,\tau,c), \tau, c) \qquad \text{und} \\
& S_{1,1}(\xi, \vartheta(\xi,\tau,c), \tau, c) = \left(-\partial_\vartheta g(\xi, \vartheta(\xi,\tau,c), \tau, c)\right)^1,
\end{aligned}
$$

was (2.58) im Fall $j = 1$ bedeutet.

Es sei nun (2.57) und (2.58) für ein $j \geq 1$ erfüllt. Zur Vereinfachung sei $C_i^{i_1,i_2,j} \in \mathbb{R}^{i_1 \times i_2}$ eine beliebige $i_1 \times i_2$-Matrix, deren Komponenten als Summe von Produkten mit mindestens einem Faktor der Form $\partial_{\zeta_1} \vartheta(\xi,\tau,c)$ bzw. $D_i(\vartheta)_{(\zeta_1,\zeta_2)}(\xi,\tau,c)$, $i = 1, \dots, j$ in jedem Produkt geschrieben werden können. Vergleichbare Matrizen wurden von

Govaerts [Gov97a] in seinem entsprechenden Beweis eingeführt. Durch Differentiation der ersten und letzten Zeile von (2.57) ergibt sich wegen

$$\begin{aligned}
d_{(\xi,\tau)}\partial_{\zeta_1} g(\xi,\vartheta(\xi,\tau,c),\tau,c) &= \partial_{(\xi,\tau)}\partial_{\zeta_1} g(\xi,\vartheta(\xi,\tau,c),\tau,c) + \\
&\quad + \partial_\vartheta \partial_{\zeta_1} g(\xi,\vartheta(\xi,\tau,c),\tau,c)\, \partial_{(\xi,\tau)}\vartheta(\xi,\tau,c)\,, \\
d_{(\xi,\tau)} D_j(g)_{(\zeta_1,\zeta_2)}(\xi,\vartheta(\xi,\tau,c),\tau,c) &= \partial_{(\xi,\tau)} D_j(g)_{(\zeta_1,\zeta_2)}(\xi,\vartheta(\xi,\tau,c),\tau,c) + \\
&\quad + \partial_\vartheta D_j(g)_{(\zeta_1,\zeta_2)}(\xi,\vartheta(\xi,\tau,c),\tau,c)\, \partial_{(\xi,\tau)}\vartheta(\xi,\tau,c)\,, \\
d_{(\xi,\tau)}\partial_{\zeta_1}\vartheta(\xi,\tau,c) &= \partial_{(\xi,\tau)}\partial_{\zeta_1}\vartheta(\xi,\tau,c) \qquad \text{und} \\
d_{(\xi,\tau)} D_j(\vartheta)_{(\zeta_1,\zeta_2)}(\xi,\tau,c) &= \partial_{(\xi,\tau)} D_j(\vartheta)_{(\zeta_1,\zeta_2)}(\xi,\tau,c)
\end{aligned}$$

die Gleichung

$$\begin{aligned}
&\begin{bmatrix} \partial^2_{\zeta_1} g & \partial_{\zeta_2}\partial_{\zeta_1} g \\ \partial_{\zeta_1} D_j(g)_{(\zeta_1,\zeta_2)} & \partial_{\zeta_2} D_j(g)_{(\zeta_1,\zeta_2)} \end{bmatrix} \\
&= \begin{bmatrix} S_{0,0}\partial^2_{\zeta_1}\vartheta & S_{0,0}\partial_{\zeta_2}\partial_{\zeta_1}\vartheta \\ S_{j,0}\partial^2_{\zeta_1}\vartheta + \sum_{i=1}^{j} S_{j,i}\partial_{\zeta_1} D_i(\vartheta)_{(\zeta_1,\zeta_2)} & S_{j,0}\partial_{\zeta_2}\partial_{\zeta_1}\vartheta + \sum_{i=1}^{j} S_{j,i}\partial_{\zeta_2} D_i(\vartheta)_{(\zeta_1,\zeta_2)} \end{bmatrix} + C_1^{2,2,j}
\end{aligned}$$

und damit

$$\begin{aligned}
D_{j+1}(g)_{(\zeta_1,\zeta_2)} &= \det \begin{bmatrix} \partial^2_{\zeta_1} g & \partial_{\zeta_2}\partial_{\zeta_1} g \\ \partial_{\zeta_1} D_j(g)_{(\zeta_1,\zeta_2)} & \partial_{\zeta_2} D_j(g)_{(\zeta_1,\zeta_2)} \end{bmatrix} \\
&= \det \begin{bmatrix} S_{0,0}\partial^2_{\zeta_1}\vartheta & S_{0,0}\partial_{\zeta_2}\partial_{\zeta_1}\vartheta \\ S_{j,j}\partial_{\zeta_1} D_j(\vartheta)_{(\zeta_1,\zeta_2)} & S_{j,j}\partial_{\zeta_2} D_j(\vartheta)_{(\zeta_1,\zeta_2)} \end{bmatrix} + C_2^{1,1,j} \\
&= S_{0,0} S_{j,j} D_{j+1}(\vartheta)_{(\zeta_1,\zeta_2)} + C_2^{1,1,j}.
\end{aligned}$$

Somit gilt (2.57) auch für $j+1$. Weiterhin ist wegen

$$\begin{aligned}
S_{j+1,j+1} = S_{0,0} S_{j,j} &= -\partial_\vartheta g(\xi,\vartheta(\xi,\tau,c),\tau,c)\ (-\partial_\vartheta g(\xi,\vartheta(\xi,\tau,c),\tau,c))^j \\
&= (-\partial_\vartheta g(\xi,\vartheta(\xi,\tau,c),\tau,c))^{j+1}
\end{aligned}$$

auch (2.58) für $j+1$ erfüllt. Deshalb gelten (2.57) und (2.58) für alle $j = 1, 2, \ldots$.

Wegen (2.51) ergibt sich für $(\xi,\tau,c) = (0,\tau^*,0)$ induktiv, daß aus

$$\partial_{\zeta_1} g(0,\vartheta^*,\tau^*,0) = 0 \quad \wedge \quad D_i(g)_{(\zeta_1,\zeta_2)}(0,\vartheta^*,\tau^*,0) = 0 \quad \forall i, 1 \le i \le j$$

$$\partial_{\zeta_1}\vartheta(0,\tau^*,0) = 0 \quad \wedge \quad D_i(\vartheta)_{(\zeta_1,\zeta_2)}(0,\tau^*,0) = 0 \quad \forall i, 1 \le i \le j$$

folgt, woraus sich

$$\left(D_{j+1}(g)_{(\zeta_1,\zeta_2)}(0,\vartheta^*,\tau^*,0) = 0 \quad \Longleftrightarrow \quad D_{j+1}(\vartheta)_{(\zeta_1,\zeta_2)}(0,\tau^*,0) = 0\right)$$

ergibt.

g (R4) ϑ: Auch hier seien $q = 1$, $n + p + k + l \ge m + 2$ und damit $g, \vartheta \in \mathbb{R}^1$ und $(\xi,\tau) \in \mathbb{R}^{n+p+k+l-m}$. Deshalb enthält (ξ,τ) mindestens zwei Komponenten. Zwei dieser Komponenten seien $\zeta \in \mathbb{R}^2$. Aus

$$\partial_\zeta g(0,\vartheta^*,\tau^*,0) = 0 \quad \text{und} \quad \partial^2_\zeta g(0,\vartheta^*,\tau^*,0) = 0$$

folgt dann wegen $g\,(R1)\,\vartheta$ und $g\,(R2)\,\vartheta$

$$\partial_\zeta \vartheta(0,\tau^*,0) = 0 \quad \text{und} \quad \partial_\zeta^2 \vartheta(0,\tau^*,0) = 0\,.$$

Aus der dritten Ableitung der zweiten Gleichung in (2.52) bezüglich ζ ergibt sich im singulären Punkt bei Unterdrückung aller verschwindenden Terme

$$\partial_\zeta^3 g(0,\vartheta^*,\tau^*,0) = -\partial_\vartheta g(0,\vartheta^*,\tau^*,0)\partial_\zeta^3 \vartheta(0,\tau^*,0)\,.$$

$D_\infty(g)_\zeta$ ist eine Summe von Produkten von vier Ableitungen dritter Ordnung der Funktion g bezüglich der Komponenten von ζ. Deshalb gilt

$$D_\infty(g)_\zeta(0,\vartheta^*,\tau^*,0) = (\partial_\vartheta g(0,\vartheta^*,\tau^*,0))^4\, D_\infty(\vartheta)_\zeta(0,\tau^*,0)$$

und wegen (2.51)

$$\operatorname{sgn} D_\infty(g)_\zeta(0,\vartheta^*,\tau^*,0) = \operatorname{sgn} D_\infty(\vartheta)_\zeta(0,\tau^*,0)\,.$$

$g\,(R5)\,\vartheta$: Es seien $q = 2$, $n = m$, $p + k + l \geq 2$ und damit $g, \vartheta, \xi \in \mathbb{R}^2$. Wegen $\partial_\xi g(0,\vartheta^*,\tau^*,0) = 0$ und $g\,(R1)\,\vartheta$ ergibt sich

$$\partial_\xi \vartheta(0,\tau^*,0) = 0\,.$$

Aus der zweiten Ableitung der zweiten Gleichung in (2.52) bezüglich ξ im singulären Punkt $(0,\tau^*,0)$ erhält man bei Unterdrückung aller verschwindenden Terme

$$\partial_\xi^2 g(0,\vartheta^*,\tau^*,0) = -\partial_\vartheta g(0,\vartheta^*,\tau^*,0)\partial_\xi^2 \vartheta(0,\tau^*,0)$$

und damit

$$\det\left(\partial_\xi^2 g(0,\vartheta^*,\tau^*,0)\begin{bmatrix}\varrho\\ \sigma\end{bmatrix}\right) = \det\left(\partial_\vartheta g(0,\vartheta^*,\tau^*,0)\right)\det\left(\partial_\xi^2\vartheta(0,\tau^*,0)\begin{bmatrix}\varrho\\ \sigma\end{bmatrix}\right)\,.$$

Folglich besitzen

$$0 = \det\left(\partial_\xi^2 g(0,\vartheta^*,\tau^*,0)\begin{bmatrix}\varrho\\ \sigma\end{bmatrix}\right) \quad \text{und} \quad 0 = \det\left(\partial_\xi^2\vartheta(0,\tau^*,0)\begin{bmatrix}\varrho\\ \sigma\end{bmatrix}\right)$$

die selbe Lösungsmannigfaltigkeit $\begin{bmatrix}\varrho\\ \sigma\end{bmatrix}$, woraus

$$\operatorname{sgn} D_{2,2}(g)(0,\vartheta^*,\tau^*,0) = \operatorname{sgn} D_{2,2}(\vartheta)(0,\tau^*,0)$$

folgt.

$g\,(R6)\,\vartheta$: Es seien $q = 2$, $n = m$, $k+l \geq 2$ und damit $g, \vartheta, \xi \in \mathbb{R}^2$. Weiterhin sei $\lambda \in \mathbb{R}$ keine Komponente von ϑ. Wegen $\partial_\xi g(0,\vartheta^*,\tau^*,0) = 0$ und $g\,(R1)\,\vartheta$ ergibt sich ebenfalls $\partial_\xi\vartheta(0,\tau^*,0) = 0$ und

$$\partial_\xi^2 g(0,\vartheta^*,\tau^*,0)\begin{bmatrix}\varrho\\ \sigma\end{bmatrix}\begin{bmatrix}\varrho\\ \sigma\end{bmatrix} = -\partial_\vartheta g(0,\vartheta^*,\tau^*,0)\partial_\xi^2\vartheta(0,\tau^*,0)\begin{bmatrix}\varrho\\ \sigma\end{bmatrix}\begin{bmatrix}\varrho\\ \sigma\end{bmatrix} \qquad \text{und}$$

$$\partial_\lambda g(0,\vartheta^*,\tau^*,0) = -\partial_\vartheta g(0,\vartheta^*,\tau^*,0)\partial_\lambda\vartheta(0,\tau^*,0)\,.$$

Wegen $\partial_\xi^2 g(0,\vartheta^*,\tau^*,0)\begin{bmatrix}\varrho_i\\ \sigma_i\end{bmatrix}\begin{bmatrix}\varrho_i\\ \sigma_i\end{bmatrix} \neq 0$, $i=1,2$, existieren Zahlen t_1 und t_2 mit

$$\partial_\lambda g(0,\vartheta^*,\tau^*,0) = \partial_\xi^2 g(0,\vartheta^*,\tau^*,0)\begin{bmatrix}\varrho_1\\ \sigma_1\end{bmatrix}\begin{bmatrix}\varrho_1\\ \sigma_1\end{bmatrix} t_1 + \partial_\xi^2 g(0,\vartheta^*,\tau^*,0)\begin{bmatrix}\varrho_2\\ \sigma_2\end{bmatrix}\begin{bmatrix}\varrho_2\\ \sigma_2\end{bmatrix} t_2 .$$

Daraus folgt mit (2.51)

$$\begin{aligned}\partial_\lambda\vartheta(0,\tau^*,0) &= -\partial_\vartheta g(0,\vartheta^*,\tau^*,0)^{-1}\partial_\lambda g(0,\vartheta^*,\tau^*,0)\\ &= -\partial_\vartheta g(0,\vartheta^*,\tau^*,0)^{-1}\left(\partial_\xi^2 g(0,\vartheta^*,\tau^*,0)\begin{bmatrix}\varrho_1\\ \sigma_1\end{bmatrix}\begin{bmatrix}\varrho_1\\ \sigma_1\end{bmatrix} t_1 + \right.\\ &\qquad \left. + \partial_\xi^2 g(0,\vartheta^*,\tau^*,0)\begin{bmatrix}\varrho_2\\ \sigma_2\end{bmatrix}\begin{bmatrix}\varrho_2\\ \sigma_2\end{bmatrix} t_2\right)\\ &= \partial_\xi^2\vartheta(0,\tau^*,0)\begin{bmatrix}\varrho_1\\ \sigma_1\end{bmatrix}\begin{bmatrix}\varrho_1\\ \sigma_1\end{bmatrix} t_1 + \partial_\xi^2\vartheta(0,\tau^*,0)\begin{bmatrix}\varrho_2\\ \sigma_2\end{bmatrix}\begin{bmatrix}\varrho_2\\ \sigma_2\end{bmatrix} t_2 .\end{aligned}$$

Damit ist die Ableitungsäquivalenz der Funktionen g aus (2.50) und ϑ aus (2.52) gezeigt. □

Bemerkung 2.38. Wegen $g\,(R5)\,\vartheta$ folgt aus $D_{2,2}(g)(0,\vartheta^*,\tau^*,0) > 0$ auch $D_{2,2}(\vartheta)(0,\tau^*,0) > 0$. Aus dem Beweis von $g\,(R5)\,\vartheta$ folgt weiterhin, daß die in $g\,(R6)\,\vartheta$ verwendete Lösungsmenge $\{t\begin{bmatrix}\varrho_i\\ \sigma_i\end{bmatrix} : \varrho_i,\sigma_i \text{ fest}, t\in\mathbb{R}, i=1,2\}$ von $0=\det\left(\partial_\xi^2 g(0,\vartheta^*,\tau^*,0)\begin{bmatrix}\varrho\\ \sigma\end{bmatrix}\right)$ gerade die Lösungsmenge von $0=\det\left(\partial_\xi^2\vartheta(0,\tau^*,0)\begin{bmatrix}\varrho\\ \sigma\end{bmatrix}\right)$ ist.

2.3 Klassifikation singulärer Punkte

Mit Hilfe der in (2.55) definierten Funktion $\boldsymbol{\vartheta}$ lassen sich singuläre Punkte klassifizieren. Diese Klassifikation kann dann zur Konstruktion erweiterter Systeme genutzt werden. Dabei können die Klassifikation von Golubitsky/Schaeffer [GS85] für die Funktion g aus Definition 2.8 und die Klassifikationen von Govaerts [Gov95] übertragen werden. Dies ist möglich, da die Funktionen g aus (2.50) und $\boldsymbol{\vartheta}$ aus (2.55) ableitungsäquivalent sind, vergleiche Bemerkung 2.33 (ii) und zum Nachweis die Sätze 2.35 und 2.37.

Für die Klassifikation betrachteten Golubitsky/Schaeffer [GS85] Keime, im Englischen germs. Keime sind Äquivalenzklassen von Funktionen, die in einer hinreichend kleinen Umgebung des singulären Punktes übereinstimmen. Dann definierten Golubitsky und Schaeffer für die Keime die starke Äquivalenzrelation, vergleiche Bemerkung 2.23. Insbesondere sind die verschiedenen Ljapunov-Schmidt-reduzierten Funktionen einer Funktion in der Umgebung eines singulären Punktes, als Keime betrachtet, bis auf gewisse Vorzeichen stark äquivalent. Für die durch die starke Äquivalenz definierten Äquivalenzklassen von Keimen bzw. für die dadurch charakterisierten Typen singulärer Punkte leiteten Golubitsky/Schaeffer [GS85] definierende Gleichungen, Nichtentartungsbedingungen und Normalformen zur Charakterisierung in den einfachsten Fällen her. Später wurden die definierenden Gleichungen z. B. von Govaerts [Gov97a] in erweiterten Systemen zur Berechnung der singulären Punkte verwendet. Der Nachweis der Regularität der Jacobimatrix des erweiterten Systems im singulären Punkt erfolgte dabei mittels der Nichtentartungsbedingungen. Deshalb ist

es für die numerische Berechnung nicht notwendig, daß die definierenden Gleichungen und Nichtentartungsbedingungen die Äquivalenzklassen der verallgemeinerten Kontaktäquivalenz eindeutig charakterisieren. Stattdessen sollten die verschiedenen Ljapunov-Schmidt-reduzierten Funktionen der gleichen Funktion in einer Umgebung des gleichen singulären Punktes gleich charakterisiert werden. Weiterhin sollte die Ausgangsgleichung (2.1) und einige der definierenden Gleichungen ein erweitertes System bezüglich des betrachteten singulären Punktes bilden. Dabei soll die Regularität der Jacobimatrix im singulären Punkt durch die Nichtentartungsbedingungen und eine Bedingung an $\alpha \in \mathbb{R}^k$ gesichert werden. Die Bedingung an α bedeutet dabei gerade bei Verwendung aller definierenden Gleichungen, daß für die Funktion g aus (2.33) $g(\xi, \lambda, \alpha)$ eine universelle Entfaltung von $g(\xi, \lambda, \alpha^*)$ ist, vergleiche Govaerts [Gov97a]. Somit können mehrere singuläre Punkte aus der Klassifikation von Golubitsky/Schaeffer [GS85] bzw. Govaerts [Gov95] zu einem singulären Punkt zusammengefaßt werden. Beispielsweise werden durch die Definition 2.32 bestimmte Vorzeichen in den Nichtentartungsbedingungen nicht beibehalten. Dies wird nicht verlangt, da sonst die Funktionen g aus (2.50) und ϑ aus (2.52) nicht ableitungsäquivalent wären. Analog wurde auch in Definition 2.22 im Gegensatz zur starken Äquivalenz nicht gefordert, daß bestimmte Vorzeichenbedingungen erfüllt sind. Dies führt gerade dazu, daß in den Relationen $g\,(R1)\,h$, $g\,(R2)\,h$ und $g\,(R3)\,h$ nicht das gleiche Vorzeichen gefordert wird. Aus den Vorzeichen in den Nichtentartungsbedingungen können im Fall $m = n$ Aussagen über die Stabilität der invarianten Lösungsmenge von $\dot{x} = F(x, \lambda, \alpha)$ gemacht werden. Dies spielt hier jedoch keine Rolle.

Im Folgenden werden die Dimensionen, definierenden Gleichungen und Nichtentartungsbedingungen für die Klassifikation mittels $\boldsymbol{\vartheta}$, ausgehend von den Klassifikationen von Golubitsky/Schaeffer [GS85], Jepson/Spence [JS84] und Govaerts [Gov95] und [Gov97a] angegeben. Damit ist für $T_2 = 0$ bei der Definition von $\boldsymbol{T}$ gemäß (2.54) auch eine Klassifikation mittels ϑ gegeben.

2.3.1 Der Fall $m = n$, $q = 1$, $p = 1$ und $\vartheta = \lambda$

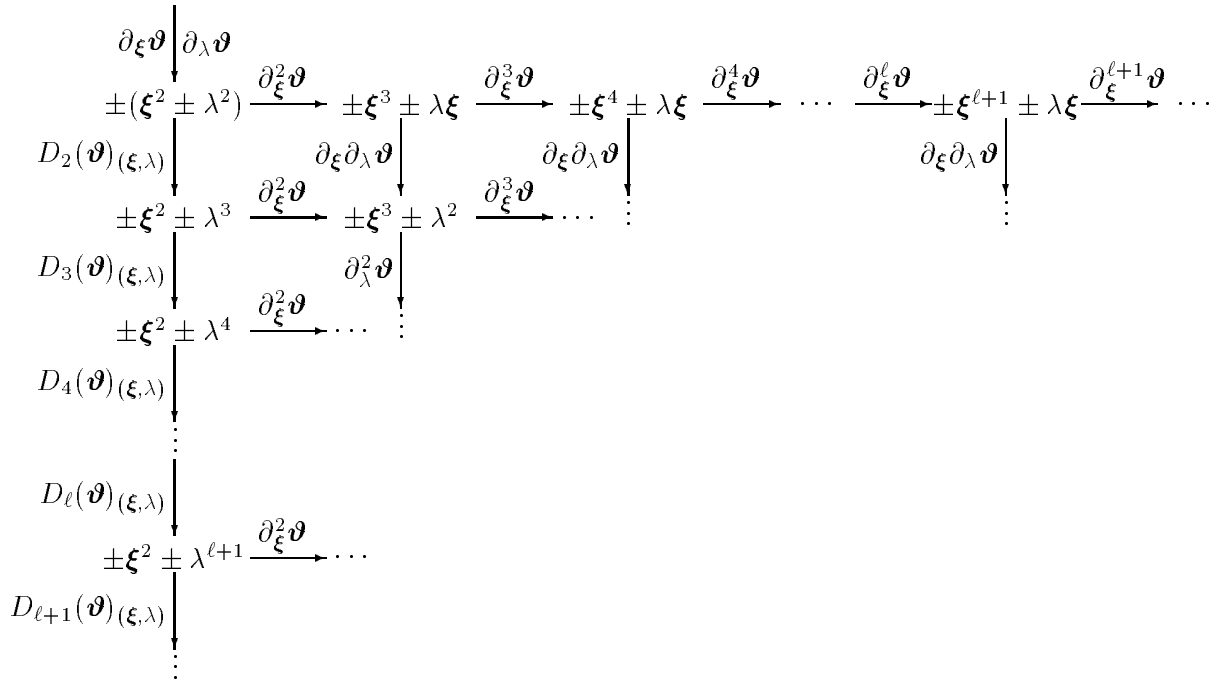
Rückkehrpunkte der Vielfachheit $\ell \geq 1$		
	Klassifikation wie in [GS85]	Neue Klassifikation
Normalform	$g = \pm\xi^{\ell+1} \pm \lambda$	$\boldsymbol{\vartheta} = \pm\boldsymbol{\xi}^{\ell+1}$
Definierende Gleichungen	$\partial_\xi g = \partial_\xi^2 g = \ldots = \partial_\xi^\ell g = 0$	$\partial_{\boldsymbol{\xi}} \boldsymbol{\vartheta} = \partial_{\boldsymbol{\xi}}^2 \boldsymbol{\vartheta} = \ldots = \partial_{\boldsymbol{\xi}}^\ell \boldsymbol{\vartheta} = 0$
Nichtentartungsbedingungen	$\partial_\xi^{\ell+1} g \neq 0\,,\ \partial_\lambda g \neq 0$	$\partial_{\boldsymbol{\xi}}^{\ell+1} \boldsymbol{\vartheta} \neq 0$

2.3.2 Der Fall $m = n$, $q = 1$ und $p = 0$

Dieser Fall wurde, ausgehend von Beyn [Bey84], von Govaerts [Gov95] und [Gov97b] betrachtet. Es ergibt sich die gleiche neue Klassifikation wie in obiger Tabelle, wobei jedoch ϑ eine Komponente von α bzw. gleich μ ist.

2.3.3 Der Fall $m = n$, $q = 1$, $p = 1$ und $\vartheta \neq \lambda$

Hier ergibt sich folgendes Schema, das die entsprechenden Teile der Schemen von Golubitsky/Schaeffer [GS85, Kapitel IV, Tabelle 2.4], Jepson/Spence [JS84, S. 199], Schnabel [Sch94] und Govaerts [Gov97a, Fig. 1] zusammenfaßt. Dabei bedeuten wie bei Jepson/Spence [JS84, S. 199] die Knotenbezeichnungen die Normalformen, während die Kantenbezeichnungen die definierenden Gleichungen bzw. die Nichtentartungsbedingungen charakterisieren.



Beachte, daß die Normalformen $g_0 := \pm\xi^{\ell+1} \pm \lambda^2$ und $g_1 := \pm\xi^{\ell+1} \pm \xi^{\ell_1}\lambda \pm \lambda^2$, $1 < \ell_1 < \ell$ i. allg. nicht äquivalent bezüglich der verallgemeinerten Kontaktäquivalenz sind, obwohl beide Normalformen die gleichen definierenden Gleichungen

$$\partial_\xi\vartheta = \partial_\xi^2\vartheta = \ldots = \partial_\xi^\ell\vartheta = \partial_\lambda\vartheta = \partial_\xi\partial_\lambda\vartheta = 0$$

und Nichtentartungsbedingungen

$$\partial_\xi^{\ell+1}\vartheta \neq 0, \;\; \partial_\lambda^2\vartheta \neq 0$$

erfüllen. Somit können auch hier all diese singulären Punkte zu jeweils einem Typ singulärer Punkte zusammengefaßt werden. Ähnliche Überlegungen gelten auch für die Klassifikation der singulären Punkte, für die $m = n$, $p = q = 1$, $\partial_{(\xi,\lambda)}\vartheta = 0$ und $\partial^2_{(\xi,\lambda)}\vartheta = 0$ gilt.

Es gibt einige andere singuläre Punkte mit Kodimension ≤ 3 im Fall $m = n$ und $p = 1$ und mit Kodimension ≤ 4 im Fall $m \in \{n, n-1\}$, $p = 0$, wie Govaerts in [Gov95], vergleiche [Gov97b], ausführte. Auch hier bleiben einige Vorzeichen nicht erhalten, weshalb bestimmte singuläre Punkte zu einem Typ zusammengefaßt werden. In den Nichtentartungsbedingungen werden dabei die Relationen „$<$“ und „$>$“ zur Relation „$\neq$“ vereinigt. Bei den Kantenbezeichnungen bedeutet „und“, daß in den definierenden Gleichungen beide Ausdrücke gleich Null sind, während in den Nichtentartungsbedingungen mindestens ein Ausdruck nicht verschwindet.

2.3.4 Der Fall $m = n - 1$, $q = 1$ und $p = 0$

Die singulären Punkte mit Kodimension ≤ 4 sind

$$
\begin{array}{l}
\Big\downarrow \partial_\xi \vartheta \\
\pm(\xi_1^2 \pm \xi_2^2) \\
\Big\downarrow D_2(\vartheta)_{(\xi_1,\xi_2)} = D_2(\vartheta)_{(\xi_2,\xi_1)} \\
\pm\xi_1^3 \pm \xi_2^2 \xrightarrow{\partial_\xi^2 \vartheta} \left\{ \begin{array}{ll} \pm\xi_1^3 \pm \xi_2^3 & \text{falls } D_\infty(\vartheta)_{(\xi_1,\xi_2)} > 0 \\ \pm(\xi_1^3 - \xi_1\xi_2^2) & \text{falls } D_\infty(\vartheta)_{(\xi_1,\xi_2)} < 0 \end{array} \right\} \xrightarrow{D_\infty(\vartheta)_{(\xi_1,\xi_2)}} \cdots \\
\Big\downarrow D_3(\vartheta)_{(\xi_1,\xi_2)} \text{ und } D_3(\vartheta)_{(\xi_2,\xi_1)} \\
\pm(\xi_1^4 \pm \xi_2^2) \\
\Big\downarrow D_4(\vartheta)_{(\xi_1,\xi_2)} \text{ und } D_4(\vartheta)_{(\xi_2,\xi_1)} \\
\pm\xi_1^5 \pm \xi_2^2 \\
\Big\downarrow D_5(\vartheta)_{(\xi_1,\xi_2)} \text{ und } D_5(\vartheta)_{(\xi_2,\xi_1)} \\
\vdots
\end{array}
$$

2.3.5 Der Fall $m = n$, $q = 2$ und $p = 0$

In diesem Fall ist g, ϑ, $\xi \in \mathbb{R}^2$ und

$$\partial_\xi \vartheta(0,\tau^*,0) = 0 : \text{Fall } D_{2,2}(\vartheta)(0,\tau^*,0) > 0 : \text{Normalform z. B. } \begin{bmatrix} \xi_1^2 + \xi_2^2 \\ \xi_1\xi_2 \end{bmatrix}$$

$$\text{Fall } D_{2,2}(\vartheta)(0,\tau^*,0) < 0 : \text{Normalform z. B. } \begin{bmatrix} \xi_1^2 - \xi_2^2 \\ \xi_1\xi_2 \end{bmatrix}$$

2.3.6 Der Fall $m = n$, $q = 2$, $p = 1$ und λ ist keine Komponente von ϑ

In diesem Fall ist ebenfalls g, ϑ, $\xi \in \mathbb{R}^2$ und $\partial_\xi \vartheta(0,\tau^*,0) = 0$. Aus der Klassifikation von Golubitsky/Schaeffer [GS85], vergleiche Govaerts [Gov95], ergibt sich
$\partial_\lambda \vartheta(0,\tau^*,0) \neq 0 \in \mathbb{R}^2$ und entweder
$D_{2,2}(\vartheta)(0,\tau^*,0) < 0$ mit der Normalform $\begin{bmatrix} \xi_1^2 - \xi_2^2 + \lambda \\ 2\xi_1\xi_2 \end{bmatrix}$ oder
$D_{2,2}(\vartheta)(0,\tau^*,0) > 0$, wobei die Lösung von $0 = \det\left(\partial_\xi^2 \vartheta(0,\tau^*,0) \begin{bmatrix} \varrho \\ \sigma \end{bmatrix}\right)$ aus zwei sich schneidenden Geraden

$$\left\{ t \begin{bmatrix} \varrho_i \\ \sigma_i \end{bmatrix} : i = 1,2, \quad \varrho_i \text{ und } \sigma_i \text{ fest}, \quad t \in \mathbb{R} \right\} \tag{2.59}$$

besteht. In diesem Fall wird die Forderung $\partial_\lambda \vartheta(0,\tau^*,0) \neq 0 \in \mathbb{R}^2$ weiter verschärft und in mehrere Fälle aufgeteilt. Dazu werden die beiden Strahlen

$$\partial_\xi^2 \vartheta(0,\tau^*,0) \left[t \begin{bmatrix} \varrho_i \\ \sigma_i \end{bmatrix} \right] \left[t \begin{bmatrix} \varrho_i \\ \sigma_i \end{bmatrix} \right] = t^2 \partial_\xi^2 \vartheta(0,\tau^*,0) \begin{bmatrix} \varrho_i \\ \sigma_i \end{bmatrix} \begin{bmatrix} \varrho_i \\ \sigma_i \end{bmatrix}, \quad t \in \mathbb{R}$$

auf der Lösungsmenge (2.59) betrachtet. Es wird zusätzlich gefordert, daß $\partial_\lambda \vartheta(0,\tau^*,0)$ nicht parallel zu einer diesen beiden Strahlen ist. Dann werden drei Fälle unterschieden:

Fall 1: $\partial_\lambda \vartheta(0,\tau^*,0) \in \text{cone}\left\{ \partial_\xi^2 \vartheta(0,\tau^*,0) \begin{bmatrix} \varrho_1 \\ \sigma_1 \end{bmatrix} \begin{bmatrix} \varrho_1 \\ \sigma_1 \end{bmatrix}, \partial_\xi^2 \vartheta(0,\tau^*,0) \begin{bmatrix} \varrho_2 \\ \sigma_2 \end{bmatrix} \begin{bmatrix} \varrho_2 \\ \sigma_2 \end{bmatrix} \right\}$:

$$\text{Normalform } \begin{bmatrix} \xi_1^2 + \lambda \\ \xi_2^2 + \lambda \end{bmatrix}$$

Fall 2: $\partial_\lambda \boldsymbol{\vartheta}(0,\tau^*,0) \in \operatorname{cone}\left\{ -\partial^2_{\boldsymbol{\xi}} \boldsymbol{\vartheta}(0,\tau^*,0) \begin{bmatrix} \varrho_1 \\ \sigma_1 \end{bmatrix} \begin{bmatrix} \varrho_1 \\ \sigma_1 \end{bmatrix}, -\partial^2_{\boldsymbol{\xi}} \boldsymbol{\vartheta}(0,\tau^*,0) \begin{bmatrix} \varrho_2 \\ \sigma_2 \end{bmatrix} \begin{bmatrix} \varrho_2 \\ \sigma_2 \end{bmatrix} \right\}$:

$$\text{Normalform} \begin{bmatrix} \boldsymbol{\xi}_1^2 - \lambda \\ \boldsymbol{\xi}_2^2 - \lambda \end{bmatrix}$$

Fall 3: nicht Fall 1 und nicht Fall 2, d. h.

$$\partial_\lambda \boldsymbol{\vartheta}(0,\tau^*,0) \in \operatorname{cone}\left\{ \partial^2_{\boldsymbol{\xi}} \boldsymbol{\vartheta}(0,\tau^*,0) \begin{bmatrix} \varrho_1 \\ \sigma_1 \end{bmatrix} \begin{bmatrix} \varrho_1 \\ \sigma_1 \end{bmatrix}, -\partial^2_{\boldsymbol{\xi}} \boldsymbol{\vartheta}(0,\tau^*,0) \begin{bmatrix} \varrho_2 \\ \sigma_2 \end{bmatrix} \begin{bmatrix} \varrho_2 \\ \sigma_2 \end{bmatrix} \right\} \cup$$

$$\operatorname{cone}\left\{ -\partial^2_{\boldsymbol{\xi}} \boldsymbol{\vartheta}(0,\tau^*,0) \begin{bmatrix} \varrho_1 \\ \sigma_1 \end{bmatrix} \begin{bmatrix} \varrho_1 \\ \sigma_1 \end{bmatrix}, \partial^2_{\boldsymbol{\xi}} \boldsymbol{\vartheta}(0,\tau^*,0) \begin{bmatrix} \varrho_2 \\ \sigma_2 \end{bmatrix} \begin{bmatrix} \varrho_2 \\ \sigma_2 \end{bmatrix} \right\} :$$

$$\text{Normalform} \begin{bmatrix} \boldsymbol{\xi}_1^2 + \lambda \\ \boldsymbol{\xi}_2^2 - \lambda \end{bmatrix}$$

2.4 Charakterisierende Gleichungen im erweiterten System

2.4.1 Grundlegende Bemerkungen

In diesem Abschnitt soll es um die Herleitung erweiterter Systeme für eine ganze Reihe der im Abschnitt 2.3 beschriebenen singulären Punkte gehen. Wie bereits erwähnt, werden dabei die definierenden Gleichungen und Nichtentartungsbedingungen eine besondere Rolle spielen. Im Gegensatz zu den aus der Literatur bekannten erweiterten Systemen, die sich aus einer Klassifikation singulärer Punkte ergeben, vergleiche z. B. Jepson/Spence [JS84], Janovský [Jan88] und Govaerts [Gov97a], wird dabei einerseits die Funktion $\boldsymbol{\vartheta}$ und nicht die Funktion g verwendet. Für Spezialfälle erhält man gerade, bis auf das Vorzeichen, die erweiterten Systeme aus [PS81], [Pön87], [Pön90] und [PSS99]. Andererseits wird hier nur angegeben, welche den singulären Punkt charakterisierenden Gleichungen im erweiterten System notwendig sind. Dies führt in den meisten Fällen dazu, daß keine Anforderungen an die Dimension von α gestellt werden. In der Literatur werden dagegen i. allg. erweiterte Systeme betrachtet, bei denen die singulären Punkte stabil gegenüber Störungen von F sind. Dies bedeutet, daß bei Funktionen, die sich von F nur durch kleine Störungen unterscheiden, in einer Umgebung von $(x^*, \lambda^*, \alpha^*)$ ein singulärer Punkt der gleichen Klasse liegt. Daraus ergibt sich eine minimale Dimension von α für diese Klasse singulärer Punkte. Weiterhin sind alle definierenden Gleichungen in den charakterisierenden Gleichungen zu verwenden. Bei den hier vorgeschlagenen Systemen wird dagegen nur gesichert, daß sie erweiterte Systeme im Sinne der Definition 2.3 bezüglich einer gegebenen Klasse singulärer Punkte sind. Ob ein regulärer Lösungspunkt des erweiterten Systems (2.7) tatsächlich einen singulären Punkt von F der entsprechenden Klasse enthält, kann nur durch Überprüfung der im erweiterten System nicht verwendeten definierenden Gleichungen erfolgen. Diese Idee ist eine Fortführung und Verallgemeinerung der entsprechenden Idee aus [PSS99].

Die erweiterten Systeme (2.10) und im Spezialfall $l = 0$ (2.8) lassen sich unter Beachtung von (2.48) auch als

$$G(y) := \begin{bmatrix} \boldsymbol{F}(x, \vartheta, \tau) \\ \boldsymbol{f}(x, \vartheta, \tau) \end{bmatrix} = 0 \in \mathbb{R}^{n+p+k+l} \tag{2.60}$$

schreiben. Für den singulären Punkt $(x^*, \vartheta^*, \tau^*)$ sind die ersten m Gleichungen stets erfüllt. Die übrigen Gleichungen können z. B. aus den definierenden Gleichungen ausgewählt werden, da sie die gleiche Form haben. Wie aus dem Abschnitt 2.3 zu erkennen ist, besteht die linke Seite der definierenden Gleichungen nur aus Ableitungen von $\boldsymbol{\vartheta}$ nach $\boldsymbol{\xi}$

bzw. λ. Dabei können die benötigten i-ten Ableitungen von $\boldsymbol{\vartheta}$ implizit aus der i-ten Ableitung von (2.55) bestimmt werden. Offensichtlich kommen in dieser x^* und ϑ^* nicht vor. Wenn weiterhin $(c, \boldsymbol{\xi}) := (\boldsymbol{F}(x, \vartheta, \tau), \boldsymbol{T}(x - x^*, \vartheta - \vartheta^*))$ gesetzt wird, ist offenbar $(\boldsymbol{x}(\boldsymbol{\xi}, \tau, c), \boldsymbol{\vartheta}(\boldsymbol{\xi}, \tau, c)) = (x, \vartheta)$. Somit hängt $\boldsymbol{f}$ nur vom aktuellen Punkt (x, ϑ, τ) ab. Darüberhinaus gilt für alle $(\boldsymbol{\xi}, \tau, c)$ in einer Umgebung von $(0, \tau^*, 0)$ und für alle (x, ϑ, τ) aus der entsprechenden Umgebung von $(x^*, \vartheta^*, \tau^*)$ sogar

$$(2.61) \quad (c, \boldsymbol{\xi}) = \left(\boldsymbol{F}(x, \vartheta, \tau), \boldsymbol{T} \begin{pmatrix} x - x^* \\ \vartheta - \vartheta^* \end{pmatrix} \right) \qquad \Leftrightarrow \qquad (\boldsymbol{x}(\boldsymbol{\xi}, \tau, c), \boldsymbol{\vartheta}(\boldsymbol{\xi}, \tau, c)) = (x, \vartheta) \, .$$

Deshalb kann eine Funktion, die von dem einen Variablensatz abhängt, lokal problemlos in eine Funktion überführt werden, die von dem anderen Variablensatz abhängt. Im Folgenden wird deshalb gelegentlich von den einen zu den anderen Variablen übergegangen. Damit (2.60) ein erweitertes System ist, ist jedoch noch zu sichern, daß $\begin{bmatrix} \partial \boldsymbol{F}(x^*, \vartheta^*, \tau^*) \\ \partial \boldsymbol{f}(x^*, \vartheta^*, \tau^*) \end{bmatrix}$ regulär ist. Eine notwendige und hinreichende Bedingung dafür beschreibt die folgende Proposition, die eine Übertragung von Proposition 3.3 aus [Gov97a] ist.

Proposition 2.39.
$\begin{bmatrix} \partial \boldsymbol{F}(x^*, \vartheta^*, \tau^*) \\ \partial \boldsymbol{f}(x^*, \vartheta^*, \tau^*) \end{bmatrix}$ *ist genau dann regulär, wenn* $d_{(\boldsymbol{\xi}, \tau)} \boldsymbol{f} = \partial_{(\boldsymbol{\xi}, \tau)} \boldsymbol{f}$ *im singulären Punkt* $(0, \tau^*, 0)$ *regulär ist, vergleiche (2.61).*

Beweis. $\partial \boldsymbol{F}(x^*, \vartheta^*, \tau^*)$ ist wegen (2.48) zeilenregulär. Somit ist

$$\dim \ker \partial \boldsymbol{F}(x^*, \vartheta^*, \tau^*) = n + p + k + l - m$$

und wegen der ersten Ableitung (2.55) nach $(\boldsymbol{\xi}, \tau)$ im singulären Punkt

$$\ker \partial \boldsymbol{F}(x^*, \vartheta^*, \tau^*) = \operatorname{im} \begin{bmatrix} \partial_{\boldsymbol{\xi}} \boldsymbol{x}(0, \tau^*, 0) & \partial_\tau \boldsymbol{x}(0, \tau^*, 0) \\ \partial_{\boldsymbol{\xi}} \boldsymbol{\vartheta}(0, \tau^*, 0) & \partial_\tau \boldsymbol{\vartheta}(0, \tau^*, 0) \\ 0 & I_{p+k+l-q} \end{bmatrix} .$$

Somit ist $\begin{bmatrix} \partial \boldsymbol{F}(x^*, \vartheta^*, \tau^*) \\ \partial \boldsymbol{f}(x^*, \vartheta^*, \tau^*) \end{bmatrix}$ genau dann regulär, wenn $\partial \boldsymbol{f}(x^*, \vartheta^*, \tau^*)$, angewendet auf diese Nullraumbasis, regulär ist. Dies ist aber gerade $d_{(\boldsymbol{\xi}, \tau)} \boldsymbol{f} = \partial_{(\boldsymbol{\xi}, \tau)} \boldsymbol{f}$ im singulären Punkt $(0, \tau^*, 0)$, woraus die Aussage folgt. □

Wie im Abschnitt 2.3 angegeben, wird unter Beachtung von Proposition 2.39 diese Regularität durch die Nichtentartungsbedingungen und eine Bedingung an α gesichert. Bei einfacheren singulären Punkten sind die Nichtentartungsbedingungen sehr einfach. Deren linken Seiten sind entweder gewisse Ableitungen der linken Seiten von definierenden Gleichungen oder deren Determinanten. Deshalb können in diesem Fall die definierenden Gleichungen im erweiterten System als charakterisierende Gleichungen verwendet werden. Falls die Nichtentartungsbedingungen jedoch komplizierter sind, lassen sich deren linken Seiten nicht als Ableitung der linken Seiten von definierenden Gleichungen bzw. als deren Determinanten darstellen. Deshalb können bei diesen singulären Punkten nicht direkt die definierenden Gleichungen als charakterisierende Gleichungen im erweiterten System verwendet werden. Stattdessen sind andere charakterisierende Funktionen, die linken Seiten der charakterisierenden Gleichungen $N^T_{m_{ges}, m_{ges}-m} G(y) = 0$, zu wählen. Diese sollen für

$y = y^*$ verschwinden. Weiterhin sind sie so zu wählen, daß deren Ableitung nach $\boldsymbol{\xi}$ bzw. nach $(\boldsymbol{\xi}, \lambda)$ bzw. deren Determinante die linke Seite von Nichtentartungsbedingungen ergibt. Damit lassen sich aus (2.1) und z. B. den definierenden Gleichungen erweiterte Systeme für einen gesuchten singulären Punkt konstruieren.

Die Bedingung an α ist leicht zu ermitteln, weshalb sie nicht weiter untersucht werden soll. Falls nämlich die Ableitung von $\boldsymbol{f}$ nach $\boldsymbol{\xi}$ und, falls vorhanden und keine Komponente von ϑ, nach λ im Punkt $(0, \tau^*, 0)$ spaltenregulär ist, ist die Bedingung an α, daß $\partial_{(\boldsymbol{\xi},\tau)}\boldsymbol{f}(0, \tau^*, 0)$ regulär ist. Da wegen der definierenden Gleichungen einige Elemente dieser Matrix Null sind, kann für eine Reihe singulärer Punkte dies gesichert werden, wenn eine Teilmatrix regulär ist. Damit ergibt sich

Lemma 2.40.
Die Funktion $\boldsymbol{f}$ bestehe aus Ableitungen von $\boldsymbol{\vartheta}$ nach $\boldsymbol{\xi}$ bzw. gegebenenfalls nach λ. Dann ist (2.60) genau dann ein erweitertes System von $\boldsymbol{F}(x, \vartheta, \tau) = 0$ bezüglich des singulären Punktes $(x^*, \vartheta^*, \tau^*)$, wenn

(2.62) *im singulären Punkt $\boldsymbol{f} = 0$ und $\partial_{(\boldsymbol{\xi},\lambda)}\boldsymbol{f}$ spaltenregulär*

und durch eine weitere Bedingung allein an α die Regularität von $\partial_{(\boldsymbol{\xi},\tau)}\boldsymbol{f}(0, \tau^*, 0)$ gewährleistet ist.

Bemerkung 2.41.

(i) Falls (2.62) erfüllt ist, hängt die Regularität von $\partial_{(\boldsymbol{\xi},\tau)}\boldsymbol{f}(0, \tau^*, 0)$ nur von der Ableitung von $\boldsymbol{f}$ nach den α-Komponenten von τ im Punkt $(0, \tau^*, 0)$ in Abhängigkeit von $\partial_{(\boldsymbol{\xi},\lambda)}\boldsymbol{f}(0, \tau^*, 0)$ ab. Dies ist die erwähnte Bedingung an α.

(ii) Für eine gegebene Klasse singulärer Punkte ist also (2.62) zu sichern, damit (2.60) unter Beachtung von Proposition 2.39 ein erweitertes System ist.

Bemerkung 2.42. In den Ausgangsvariablen ergibt sich:

(i) Wegen (2.3), (2.9) und (2.48) sind offensichtlich die Komponenten von μ im Fall $l > 0$ Komponenten von ϑ. Somit läßt sich ϑ in (θ, μ) zerlegen, wobei die Komponenten von θ Komponenten von (λ, α) sind.

(ii) Im Fall $l = 0$ ist somit $\vartheta = \theta$.

(iii) Im Fall $l = q$ ist $\vartheta = \mu$. Wegen (2.48) entspricht dann (2.37) (2.55) mit $x = \boldsymbol{x}$, $\mp D = \bar{D}$, $g = \boldsymbol{\vartheta}$, $\begin{bmatrix} T & \mp\delta \end{bmatrix} = \boldsymbol{T}$ und $\xi = \boldsymbol{\xi}$.

(iv) Es sei $l > 0$. In der i-ten Ableitung von (2.55) kommen außer der i-ten Ableitung von $\boldsymbol{\mu}$ keine weiteren Ableitungen von $\boldsymbol{\mu}$ vor. Bei der Berechnung der i-ten Ableitung von $\boldsymbol{\vartheta}$ wird die i-te Ableitung von (2.55) nach dieser i-ten Ableitung von $\boldsymbol{\vartheta}$ aufgelöst. Eventuell auftretende Ableitungen impliziter Funktionen niedrigerer Ordnung werden analog eliminiert. Somit hängt die i-te Ableitung von $\boldsymbol{\vartheta}$ und damit auch die i-te Ableitung von $\boldsymbol{\mu}$ nur von konstanten Matrizen und von Ableitungen von F ab. Bei den Ableitungen von F kann wegen (2.61) das Argument (x, λ, α) verwendet werden. Somit hängen alle Ableitungen von $\boldsymbol{\vartheta}$ nicht von μ ab. Sowohl $\boldsymbol{f}$ als auch die betrachteten Richtungsableitungen in Nullraumrichtung von $\partial\boldsymbol{F}(x, \vartheta, \tau)$ bestehen nur aus Ableitungen von $\boldsymbol{\vartheta}$ nach $\boldsymbol{\xi}$ bzw. τ. Sie sind damit nicht von μ abhängig. Somit kann auch im Fall $l > 0$ $\boldsymbol{f}(x, \vartheta, \tau) = f(x, \lambda, \alpha)$ gesetzt werden, woraus sich ein erweitertes System der Form (2.10) ergibt.

2.4.2 Charakterisierende Gleichungen

Im Folgenden sollen für die im Abschnitt 2.3 klassifizierten singulären Punkte angegeben werden, welche definierenden bzw. anderen Gleichungen im erweiterten System hinreichend dafür sind, daß die Ableitung von $\boldsymbol{f}$ nach $\boldsymbol{\xi}$ und gegebenenfalls nach λ spaltenregulär ist. Dies wird durch die Nichtentartungsbedingungen gesichert. Die Verwendung weiterer definierender Gleichungen im erweiterten System beeinflußt dies nicht, sondern verändert lediglich die Bedingung an α.

2.4.2.1 Der Fall $m = n$, $q = 1$, $p = 1$ und $\vartheta = \lambda$

In diesem Fall ist $l = 0$, $\partial_{(\boldsymbol{\xi},\lambda)}\boldsymbol{f} = \partial_{\boldsymbol{\xi}}\boldsymbol{f}$ und, falls vorhanden, $\alpha = \tau$.

Lemma 2.43.
Es sei $(x^*, \lambda^*, \alpha^*)$ ein Rückkehrpunkt der Vielfachheit ℓ. Wenn $\boldsymbol{f} = 0$ aus definierenden Gleichungen dieses singulären Punktes besteht, wobei eine Komponente $\partial_{\boldsymbol{\xi}}^{\ell}\vartheta = 0$ entspricht, so ist (2.62) erfüllt.

Bemerkung 2.44.

(i) $\partial_{\boldsymbol{\xi}}\partial_{\boldsymbol{\xi}}^{\ell}\vartheta(0, \tau^*, 0) = \partial_{\boldsymbol{\xi}}^{\ell+1}\vartheta(0, \tau^*, 0) \neq 0$

(ii) Die Ableitungen der übrigen eventuell vorhandenen Komponenten von $\boldsymbol{f}$ nach $\boldsymbol{\xi}$ sind dagegen im singulären Punkt wegen der definierenden Gleichungen Null.

(iii) Deshalb ist die Bedingung an ein eventuell existierendes τ, daß die Ableitungsmatrix dieser übrigen Komponenten nach τ regulär ist. Dabei ist τ als Teil des Variablensatzes $(\boldsymbol{\xi}, \tau, c)$ zu verstehen.

(iv) Wenn all dies erfüllt ist, ist $F(x, \lambda, \alpha) = 0, f(x, \lambda, \alpha) = 0$ ein minimal erweitertes System der Form (2.8).

2.4.2.2 Der Fall $m = n$, $q = 1$ und $p = 0$

Auch in diesem Fall ist $\partial_{(\boldsymbol{\xi},\lambda)}\boldsymbol{f} = \partial_{\boldsymbol{\xi}}\boldsymbol{f}$. $\alpha = \tau$ gilt nur bei $l = 1$. Sonst ist eine Komponente von α der Parameter ϑ und gehört deshalb nicht zu τ. Analog zu 2.4.2.1 gilt auch hier:

Lemma 2.45.
Es sei ϑ in einer Umgebung des singulären Punktes ableitungsäquivalent zu der Normalform $\pm\boldsymbol{\xi}^{\ell+1}$. Wenn $\boldsymbol{f} = 0$ aus definierenden Gleichungen dieses singulären Punktes besteht, wobei eine Komponente $\partial_{\boldsymbol{\xi}}^{\ell}\vartheta = 0$ entspricht, so ist (2.62) erfüllt.

Auch die Bemerkungen 2.44 (i)–(iii) gelten auch hier, während das so erhaltene erweiterte System nur im Fall $l = 0$, d. h., ϑ ist eine Komponente von α, die Form (2.8) hat. Sonst besitzt das erweiterte System die Form (2.10).

2.4.2.3 Der Fall $m = n$, $q = 1$, $p = 1$ und $\vartheta \neq \lambda$

Hier ist $(\boldsymbol{\xi}, \lambda) \in \mathbb{R}^2$.

Lemma 2.46.
Es bestehe $\boldsymbol{f} = 0$ aus definierenden Gleichungen des singulären Punktes. Bei der Normalform $\pm\boldsymbol{\xi}^2 \pm \lambda^{\ell+1}$, $\ell \geq 1$ seien $\partial_{\boldsymbol{\xi}}\vartheta$ und $D_{\ell}(\vartheta)_{(\boldsymbol{\xi},\lambda)}$ zwei Komponenten von $\boldsymbol{f}$. Bei der Normalform $\pm\boldsymbol{\xi}^{\ell+1} \pm \lambda\boldsymbol{\xi}$, $\ell \geq 2$ sei $\partial_{\boldsymbol{\xi}}\vartheta$ eine Komponente von $\boldsymbol{f}$, während eine andere $\partial_{\lambda}\vartheta$ oder $\partial_{\boldsymbol{\xi}}^{\ell}\vartheta$ ist. Bei singulären Punkten aus 2.3.3 mit den Nichtentartungsbedingungen $\partial_{\boldsymbol{\xi}}^{\ell+1}\vartheta \neq 0$, $\ell \geq 2$ und $\partial_{\lambda}^{2}\vartheta \neq 0$ seien $\partial_{\boldsymbol{\xi}}^{\ell}\vartheta$ und $\partial_{\lambda}\vartheta$ zwei Komponenten von $\boldsymbol{f}$. Dann ist (2.62) erfüllt.

Bemerkung 2.47.

(i) Die erwähnten Komponenten von $\boldsymbol{f}$ sind offensichtlich linke Seiten der definierenden Gleichungen der entsprechenden singulären Punkte. Da $\boldsymbol{f} = 0$ nur aus definierenden Gleichungen besteht, ist diese Gleichheit im singulären Punkt gegeben.

(ii) Bei der Normalform $\pm\boldsymbol{\xi}^2 \pm \lambda^{\ell+1}$, $\ell \geq 1$ ist dann (2.62) wegen der Definition von $D_{\ell+1}(\boldsymbol{\vartheta})_{(\boldsymbol{\xi},\lambda)}$ erfüllt. Im singulären Punkt sind $\partial_{(\boldsymbol{\xi},\lambda)} D_i(\boldsymbol{\vartheta})_{(\boldsymbol{\xi},\lambda)}$ für $1 \leq i < \ell$ und $\partial_{(\boldsymbol{\xi},\lambda)}\partial_{\boldsymbol{\xi}}\boldsymbol{\vartheta}$ wegen $D_{i+1}(\boldsymbol{\vartheta})_{(\boldsymbol{\xi},\lambda)} = 0$ linear abhängig.

(iii) Bei der Normalform $\pm\boldsymbol{\xi}^{\ell+1} \pm \lambda\boldsymbol{\xi}$, $\ell \geq 2$ ist $\partial_{\boldsymbol{\xi}}\partial_{\boldsymbol{\xi}}\boldsymbol{\vartheta} = \partial_{\boldsymbol{\xi}}^2\boldsymbol{\vartheta} = 0$. Deshalb ist dann (2.62) wegen $\partial_{\boldsymbol{\xi}}\partial_\lambda\boldsymbol{\vartheta} \neq 0$ bzw. $\partial_{\boldsymbol{\xi}}\partial_\lambda\boldsymbol{\vartheta} \neq 0$ und $\partial_{\boldsymbol{\xi}}\partial_{\boldsymbol{\xi}}^\ell\boldsymbol{\vartheta} = \partial_{\boldsymbol{\xi}}^{\ell+1}\boldsymbol{\vartheta} \neq 0$ erfüllt. Bei den übrigen definierenden Gleichungen $\partial_{\boldsymbol{\xi}}^i\boldsymbol{\vartheta} = 0$ mit $1 < i < \ell$ gilt wie bei $i = 1$ ebenfalls $\partial_{\boldsymbol{\xi}}\partial_{\boldsymbol{\xi}}^i\boldsymbol{\vartheta} = \partial_{\boldsymbol{\xi}}^{i+1}\boldsymbol{\vartheta} = 0$.

(iv) Offensichtlich können bei $(\alpha, \mu) \in \mathbb{R}^1$, d. h., der eine Parameter ist eindimensional, während der andere nicht existiert, als Komponenten von $\boldsymbol{f}$ für alle singulären Punkte mit der Normalform $\pm\boldsymbol{\xi}^{\ell+1} \pm \lambda\boldsymbol{\xi}$, $\ell \geq 2$, unabhängig von ℓ, $\partial_{\boldsymbol{\xi}}\boldsymbol{\vartheta}$ und $\partial_\lambda\boldsymbol{\vartheta}$ gewählt werden. Dieses erweiterte System für den einfachen Verzweigungspunkt / Einsiedlerpunkt, Normalform $\pm(\boldsymbol{\xi}^2 \pm \lambda^2)$, ist somit ein erweitertes System für all diese singulären Punkte. Diese Aussage steht bereits in [PSS97b].

(v) Bei singulären Punkten aus 2.3.3 mit den Nichtentartungsbedingungen $\partial_{\boldsymbol{\xi}}^{\ell+1}\boldsymbol{\vartheta} \neq 0$, $\ell \geq 2$ und $\partial_\lambda^2\boldsymbol{\vartheta} \neq 0$ gilt analog $\partial_{\boldsymbol{\xi}}\partial_{\boldsymbol{\xi}}^\ell\boldsymbol{\vartheta} = \partial_{\boldsymbol{\xi}}^{\ell+1}\boldsymbol{\vartheta} \neq 0$, $\partial_{\boldsymbol{\xi}}\partial_\lambda\boldsymbol{\vartheta} = 0$ und $\partial_\lambda\partial_\lambda\boldsymbol{\vartheta} \neq 0$. Daraus folgt (2.62). Auch hier gilt $\partial_{\boldsymbol{\xi}}\partial_{\boldsymbol{\xi}}^i\boldsymbol{\vartheta} = \partial_{\boldsymbol{\xi}}^{i+1}\boldsymbol{\vartheta} = 0$ für $1 \leq i < \ell$.

2.4.2.4 Der Fall $m = n - 1$, $q = 1$ und $p = 0$

In diesem Fall ist $\boldsymbol{\xi} \in \mathbb{R}^2$ und $\partial_{(\boldsymbol{\xi},\lambda)}\boldsymbol{f} = \partial_{\boldsymbol{\xi}}\boldsymbol{f}$.

Lemma 2.48.
Es sei (x^*, α^*) ein singulärer Punkt aus 2.3.4 mit den Nichtentartungsbedingungen $D_{\ell+1}(\boldsymbol{\vartheta})_{(\boldsymbol{\xi}_1,\boldsymbol{\xi}_2)} \neq 0$ bzw. $D_{\ell+1}(\boldsymbol{\vartheta})_{(\boldsymbol{\xi}_2,\boldsymbol{\xi}_1)} \neq 0$ für $1 \leq \ell \leq 4$. $\boldsymbol{f} = 0$ bestehe aus definierenden Gleichungen des singulären Punktes. Falls $\boldsymbol{f} = 0$ im Fall $\ell = 1$ gerade aus den beiden definierenden Gleichungen $\partial_{\boldsymbol{\xi}_1}\boldsymbol{\vartheta} = 0$ und $\partial_{\boldsymbol{\xi}_2}\boldsymbol{\vartheta} = 0$ besteht, ist (2.62) erfüllt. Falls in den anderen Fällen $1 < \ell \leq 4$ vier Komponenten von $\boldsymbol{f}$ $\partial_{\boldsymbol{\xi}_1}\boldsymbol{\vartheta}$, $\partial_{\boldsymbol{\xi}_2}\boldsymbol{\vartheta}$, $D_\ell(\boldsymbol{\vartheta})_{(\boldsymbol{\xi}_1,\boldsymbol{\xi}_2)}$ und $D_\ell(\boldsymbol{\vartheta})_{(\boldsymbol{\xi}_2,\boldsymbol{\xi}_1)}$ sind, ist (2.62) ebenfalls erfüllt.

Bemerkung 2.49.

(i) Diese erweiterten Systeme wurden analog zu denen für singuläre Punkte mit der Normalform $\pm\boldsymbol{\xi}^2 \pm \lambda^{\ell+1}, \ell \geq 1$ gewählt.

(ii) Analog zu jenen Punkten ist auch bei diesen erweiterten Systemen (2.62) erfüllt. Weiterhin sind auch hier im singulären Punkt $\partial_{\boldsymbol{\xi}}\partial_{\boldsymbol{\xi}_1}\boldsymbol{\vartheta}$ und $\partial_{\boldsymbol{\xi}} D_i(\boldsymbol{\vartheta})_{(\boldsymbol{\xi}_1,\boldsymbol{\xi}_2)}$ für $1 \leq i < \ell$ wegen $D_{i+1}(\boldsymbol{\vartheta})_{(\boldsymbol{\xi}_1,\boldsymbol{\xi}_2)} = 0$ linear abhängig. Analog sind im singulären Punkt $\partial_{\boldsymbol{\xi}}\partial_{\boldsymbol{\xi}_2}\boldsymbol{\vartheta}$ und $\partial_{\boldsymbol{\xi}} D_i(\boldsymbol{\vartheta})_{(\boldsymbol{\xi}_2,\boldsymbol{\xi}_1)}$ für $1 \leq i < \ell$ wegen $D_{i+1}(\boldsymbol{\vartheta})_{(\boldsymbol{\xi}_2,\boldsymbol{\xi}_1)} = 0$ linear abhängig.

(iii) Offensichtlich ist $\partial_{\boldsymbol{\xi}_1}\boldsymbol{\vartheta} = D_1(\boldsymbol{\vartheta})_{(\boldsymbol{\xi}_2,\boldsymbol{\xi}_1)}$ und $\partial_{\boldsymbol{\xi}_2}\boldsymbol{\vartheta} = D_1(\boldsymbol{\vartheta})_{(\boldsymbol{\xi}_1,\boldsymbol{\xi}_2)}$, weshalb das erweiterte System im Fall $\ell = 1$ die Übertragung der erweiterten Systeme für $\ell > 1$ ist.

(iv) Somit sind bei $\ell > 1$ $\partial_{\boldsymbol{\xi}}\partial_{\xi_1}\boldsymbol{\vartheta}$ und $\partial_{\boldsymbol{\xi}}\partial_{\xi_2}\boldsymbol{\vartheta}$ linear abhängig. Wegen $D_{\ell+1}(\boldsymbol{\vartheta})_{(\xi_1,\xi_2)} \neq 0$ oder $D_{\ell+1}(\boldsymbol{\vartheta})_{(\xi_2,\xi_1)} \neq 0$ ist mindestens einer der beiden Vektoren $\partial_{\boldsymbol{\xi}}\partial_{\xi_i}\boldsymbol{\vartheta}$, $i \in \{1,2\}$ nicht der Nullvektor. Demnach kann im erweiterten System $\partial_{\boldsymbol{\xi}_{3-i}}\boldsymbol{\vartheta} = 0$ weggelassen werden, ohne (2.62) zu verletzen.

(v) Damit das erweiterte System aus Lemma 2.48 für $\ell > 1$ ein minimal erweitertes System ist, hat α im Fall $l = 0$ mindestens die Dimension drei, im Fall $l = 1$ mindestens die Dimension zwei. Wenn $\partial_{\boldsymbol{\xi}_{3-i}}\boldsymbol{\vartheta} = 0$ im Fall $\ell > 1$ weggelassen wird, hat demnach α bei einem minimal erweiterten System im Fall $l = 0$ mindestens die Dimension zwei, im Fall $l = 1$ mindestens die Dimension eins.

(vi) Diese Dimensionsforderung an α für minimal erweiterte Systeme tritt bei den bisher behandelten singulären Punkten nicht auf. Sie ergibt sich daraus, daß für eine gegebene Klasse singulärer Punkte von zwei Ungleichungen mindestens eine erfüllt ist, aber nicht feststeht welche.

Lemma 2.50.
Es sei (x^*, α^*) ein singulärer Punkt aus 2.3.4 mit der Nichtentartungsbedingung $D_\infty(\boldsymbol{\vartheta})_{(\xi_1,\xi_2)} \neq 0$. Außer aus definierenden Gleichungen bestehe $\boldsymbol{f} = 0$ aus

$$\begin{aligned} &\partial^2_{\xi_1}\boldsymbol{\vartheta}\,\partial^3_{\xi_2}\boldsymbol{\vartheta} + \partial^2_{\xi_1}\partial_{\xi_2}\boldsymbol{\vartheta}\,\partial^2_{\xi_2}\boldsymbol{\vartheta} - 2\partial_{\xi_1}\partial_{\xi_2}\boldsymbol{\vartheta}\,\partial_{\xi_1}\partial^2_{\xi_2}\boldsymbol{\vartheta} = 0 \qquad \textit{und} \\ &\partial^2_{\xi_1}\boldsymbol{\vartheta}\,\partial_{\xi_1}\partial^2_{\xi_2}\boldsymbol{\vartheta} + \partial^3_{\xi_1}\boldsymbol{\vartheta}\,\partial^2_{\xi_2}\boldsymbol{\vartheta} - 2\partial_{\xi_1}\partial_{\xi_2}\boldsymbol{\vartheta}\,\partial^2_{\xi_1}\partial_{\xi_2}\boldsymbol{\vartheta} = 0\,. \end{aligned} \tag{2.63}$$

Dann ist (2.62) erfüllt.

Bemerkung 2.51.

(i) Da in jedem Summanden ein Faktor eine zweite Ableitung von $\boldsymbol{\vartheta}$ nach einer Komponente von $\boldsymbol{\xi}$ ist, sind diese beiden charakterisierenden Gleichungen im singulären Punkt erfüllt.

(ii) Diese beiden Gleichungen sind gerade so gewählt, daß die Determinante der Ableitungen ihrer linken Seiten nach $\boldsymbol{\xi}$ im singulären Punkt $D_\infty(\boldsymbol{\vartheta})_{(\xi_1,\xi_2)}$ ist. Deshalb ist (2.62) erfüllt. Um dies zu sichern, sind diese Gleichungen kompliziert.

(iii) Wegen $\partial^2_{\boldsymbol{\xi}}\boldsymbol{\vartheta} = 0$ im singulären Punkt, sind die Ableitungen von $\partial_{\boldsymbol{\xi}}\boldsymbol{\vartheta}$ nach $\boldsymbol{\xi}$ in diesem Punkt Null.

2.4.2.5 Der Fall $m = n$, $q = 2$ und $p = 0$

Auch hier ist $\boldsymbol{\xi} \in \mathbb{R}^2$ und $\partial_{(\boldsymbol{\xi},\lambda)}\boldsymbol{f} = \partial_{\boldsymbol{\xi}}\boldsymbol{f}$.

Lemma 2.52.
Es sei (x^*, α^*) ein singulärer Punkt aus 2.3.5. Außer aus definierenden Gleichungen bestehe $\boldsymbol{f} = 0$ aus

$$\begin{aligned} &\partial_{\xi_1}\boldsymbol{\vartheta_1}\,\partial^2_{\xi_2}\boldsymbol{\vartheta_2} + \partial_{\xi_1}\partial_{\xi_2}\boldsymbol{\vartheta_1}\,\partial_{\xi_2}\boldsymbol{\vartheta_2} - \partial_{\xi_1}\boldsymbol{\vartheta_2}\,\partial^2_{\xi_2}\boldsymbol{\vartheta_1} - \partial_{\xi_1}\partial_{\xi_2}\boldsymbol{\vartheta_2}\,\partial_{\xi_2}\boldsymbol{\vartheta_1} = 0 \qquad \textit{und} \\ &\partial_{\xi_1}\boldsymbol{\vartheta_1}\,\partial_{\xi_1}\partial_{\xi_2}\boldsymbol{\vartheta_2} + \partial^2_{\xi_1}\boldsymbol{\vartheta_1}\,\partial_{\xi_2}\boldsymbol{\vartheta_2} - \partial_{\xi_1}\boldsymbol{\vartheta_2}\,\partial_{\xi_1}\partial_{\xi_2}\boldsymbol{\vartheta_1} - \partial^2_{\xi_1}\boldsymbol{\vartheta_2}\,\partial_{\xi_2}\boldsymbol{\vartheta_1} = 0\,. \end{aligned} \tag{2.64}$$

Dann ist (2.62) erfüllt.

Bemerkung 2.53.

(i) Entsprechend dem Zusammenhang zwischen $D_\infty(\boldsymbol{\vartheta})_{(\boldsymbol{\xi}_1,\boldsymbol{\xi}_2)}$ und $D_{2,2}(\boldsymbol{\vartheta})$ wurde (2.64) analog zu (2.63) gewählt.

(ii) Gemäß dieser Analogie ist hier wegen $\partial_{\boldsymbol{\xi}}\boldsymbol{\vartheta} = 0$ und $D_{2,2}(\boldsymbol{\vartheta}) \neq 0$ im singulären Punkt (2.62) erfüllt.

2.4.2.6 Der Fall $m = n$, $q = 2$, $p = 1$ und λ ist keine Komponente von ϑ

Daß λ keine Komponente von ϑ ist, wurde gefordert, damit die Klassifikation in 2.3.6 analog zu der in Golubitsky/Schaeffer [GS85] durchgeführt werden konnte. Da jedoch $\partial_\lambda \boldsymbol{\vartheta}(0,\tau^*,0) \neq 0$ erfüllt ist, kann nach einer Umparametrisierung λ eine Komponente ϑ sein. Dann gilt wie in 2.4.2.5 $\boldsymbol{\xi} \in \mathbb{R}^2$ und $\partial_{(\boldsymbol{\xi},\lambda)}\boldsymbol{f} = \partial_{\boldsymbol{\xi}}\boldsymbol{f}$ und analog

Lemma 2.54.
Es sei (x^*,λ^*,α^*) ein singulärer Punkt aus 2.3.6. Außer aus definierenden Gleichungen bestehe $\boldsymbol{f} = 0$ aus

$$\begin{aligned}&\partial_{\boldsymbol{\xi}_1}\boldsymbol{\vartheta_1}\,\partial^2_{\boldsymbol{\xi}_2}\boldsymbol{\vartheta_2} + \partial_{\boldsymbol{\xi}_1}\partial_{\boldsymbol{\xi}_2}\boldsymbol{\vartheta_1}\,\partial_{\boldsymbol{\xi}_2}\boldsymbol{\vartheta_2} - \partial_{\boldsymbol{\xi}_1}\boldsymbol{\vartheta_2}\,\partial^2_{\boldsymbol{\xi}_2}\boldsymbol{\vartheta_1} - \partial_{\boldsymbol{\xi}_1}\partial_{\boldsymbol{\xi}_2}\boldsymbol{\vartheta_2}\,\partial_{\boldsymbol{\xi}_2}\boldsymbol{\vartheta_1} = 0 \qquad \textit{und}\\ &\partial_{\boldsymbol{\xi}_1}\boldsymbol{\vartheta_1}\,\partial_{\boldsymbol{\xi}_1}\partial_{\boldsymbol{\xi}_2}\boldsymbol{\vartheta_2} + \partial^2_{\boldsymbol{\xi}_1}\boldsymbol{\vartheta_1}\,\partial_{\boldsymbol{\xi}_2}\boldsymbol{\vartheta_2} - \partial_{\boldsymbol{\xi}_1}\boldsymbol{\vartheta_2}\,\partial_{\boldsymbol{\xi}_1}\partial_{\boldsymbol{\xi}_2}\boldsymbol{\vartheta_1} - \partial^2_{\boldsymbol{\xi}_1}\boldsymbol{\vartheta_2}\,\partial_{\boldsymbol{\xi}_2}\boldsymbol{\vartheta_1} = 0\,.\end{aligned}$$

Dann ist (2.62) erfüllt.

Bemerkung 2.55.

(i) Da das erweiterte System aus 2.4.2.5 gewählt wurde, gilt auch hier die Bemerkung 2.53.

(ii) Dieses erweiterte System ist komplizierter als das erweiterte System von Allgower/Schwetlick [AS97]. Jenes System beruht auf einer einfacheren Klassifikation. Mit jenem erweiterten System können auch singuläre Punkte mit den Normalformen $\pm(\boldsymbol{\xi}^2 \pm \lambda^2)$ und $\pm\boldsymbol{\xi}^{\ell+1} \pm \lambda\boldsymbol{\xi}$, $\ell \geq 2$ berechnet werden.

2.4.3 Vergleichbare Systeme in der Literatur

Viele erweiterte Systeme in der Literatur ergeben sich unmittelbar aus der in dieser Arbeit beschriebenen Vorgehensweise bzw. stehen in enger Beziehung mit ihr. Einige dieser erweiterten Systeme werden in diesem Abschnitt in der Reihenfolge des Erscheinens der entsprechenden Arbeiten aufgeführt; auf die unterschiedlichen Bezeichnungen in den einzelnen Arbeiten wird dabei nicht eingegangen. Insbesondere werden dabei die charakterisierenden Funktionen, die linken Seiten der charakterisierenden Gleichungen $N^T_{m_{ges},m_{ges}-m}\,G(y) = 0$, angegeben. Ein wesentlicher Teil dieser Übersicht ist eine Überarbeitung der entsprechenden Übersicht aus [Sch98], Abschnitt 7.

- Pönisch/Schwetlick [PS81] haben 1981 für den einfachen Rückkehrpunkt ($m = n$, $p = q = 1$, $k = l = 0$, $\vartheta = \lambda$) ein erweitertes System mit der charakterisierenden Funktion $\partial_{\boldsymbol{\xi}}\boldsymbol{\lambda}$ beschrieben.

- In einem 1984 erschienenen Tagungsband einer 1983 durchgeführten Tagung schlugen Jepson/Spence [JS84] für die singulären Punkte aus 2.3.1 und 2.3.3 mit der Kodimension ≤ 3 erweiterte Systeme vor. Als charakterisierende Gleichungen wurden die definierenden Gleichungen mit den Ableitungen von g aus Definition 2.13 verwendet. Speziell wurde $v_1 = w_1 = N_{n,1}$ gesetzt, was $T^T = D = N_{n,1}$ entspricht.

Diese erweiterten Systeme wurden von Jepson/Spence [JS85] im skalaren Fall, d. h. $m = n = 1$, zur Bestimmung der Gebiete im α-Raum genutzt, in denen die Lösungsmengen von (2.1) das gleiche Verzweigungsverhalten haben.

- Griewank/Reddien [GR84] stellten 1984 erweiterte Systeme für den einfachen Rückkehrpunkt ($m = n$, $p = q = 1$, $k = l = 0$) und für spezielle Verzweigungspunkte bzw. Einsiedlerpunkte ($m = n$, $p = q = l = 1$, $k = 0$) auf. Für den einfachen Rückkehrpunkt benutzten sie $\partial_\xi g$. Für die Verzweigungs-/Einsiedlerpunkte wurde $\partial_\xi g \in \mathbb{R}^2$ aus dem etwas veränderten System

$$F(x(\xi), \lambda(\xi)) \equiv Dg(\xi), \quad T(x(\xi), \lambda(\xi)) \equiv \xi, \quad (x, \lambda, g)(T(x^*, \lambda^*)) = (x^*, \lambda^*, 0)$$

mit $T \in \mathbb{R}^{2\times(n+1)}$ und $D \in \mathbb{R}^m$ verwendet. In allen Fällen wird $\delta = 0$ gesetzt. Im Fall $l = 1$ entspricht im Spezialfall $M_{2,1}^T\, T\, N_{n+1,1} = 0$ und $N_{2,1}^T\, T = N_{n+1,1}^T$ das erweiterte System gerade dem erweiterten System aus Bemerkung 2.47 (iv). Dabei ist $\vartheta = \mu$. In allen Fällen gilt

(2.65)

$$\partial_\xi g = -N_{n+1+l,1}^T\, B^{-1}\, N_{n+1+l,1+l} \qquad \text{mit}$$

$$B(x,\lambda) = \begin{bmatrix} \partial_x F(x,\lambda) & -D \\ -T & 0 \end{bmatrix} \in \mathbb{R}^{(n+1)\times(n+1)} \quad \text{mit} \quad D, T^T \in \mathbb{R}^n \qquad \text{bzw.}$$

$$B(x,\lambda) = \begin{bmatrix} \partial F(x,\lambda) & -D \\ -T & 0 \end{bmatrix} \in \mathbb{R}^{(n+2)\times(n+2)} \quad \text{mit} \quad D \in \mathbb{R}^n,\ T^T \in \mathbb{R}^{(n+1)\times 2}.$$

Für die Funktion g aus (2.37) ergibt sich dagegen

$$(2.66) \quad \partial_\xi g = +N_{n+q,q}^T\, B^{-1}\, N_{n+q,n-m+q} \quad \text{mit} \quad B(x,\lambda,\alpha) = \begin{bmatrix} \partial_x F(x,\lambda,\alpha) & \mp D \\ T & \mp\delta \end{bmatrix}.$$

Dabei wurde wegen der anderen Gestalt von B auch das andere Vorzeichen bei der Definition von $\partial_\xi g$ verwendet.

- Auf [GR84] aufbauend, beschrieben Rabier/Reddien [RR86] 1986 ein erweitertes System für spezielle singuläre Punkte mit $1 \le m \le n$, $p = k = 0$ und $q = l \ge 1$. Als charakterisierende Funktion wurde $\pm\partial_\xi g_q$ verwendet, wobei die letzte Spalte von δ Null ist, d. h., $\delta\, N_{q,1} = 0$.

- 1987 stellte Pönisch [Pön87] ein erweitertes System für den Hysteresepunkt, d. h. den zweifachen Rückkehrpunkt, ($m = n$, $p = k = q = 1$, $l = 0$, $\vartheta = \lambda$) mit den charakterisierenden Funktionen $\partial_{\boldsymbol{\xi}}\boldsymbol{\lambda}$ und $-\partial_{\boldsymbol{\xi}}^2\boldsymbol{\lambda}$ auf.

- 1988 beschrieb Janovský [Jan88] erweiterte Systeme für singuläre Punkte in Banachräumen, wobei $\partial_x F(x,\lambda,\alpha)$ ein Fredholmoperator mit Index Null ist. Im Endlichdimensionalen bedeutet dies $m = n$. Weiterhin sei $p = q = 1$ und die Kodimension bezüglich des Parameters λ sei ≤ 2. Zur Beschreibung der charakterisierenden Funktionen $\partial_\xi g$, $\partial_\xi^2 g$, $\partial_\lambda g$, $\partial_\xi^3 g$ bzw. $\beta\partial_\xi^2 g + \partial_\xi\partial_\lambda g$ und $\beta\partial_\xi\partial_\lambda g + \partial_\lambda^2 g$ verwendet er die Matrix $B(x,\lambda,\alpha)$ mit $\delta = 0$ und $T\,D = 1$. Diese Funktionen entsprechen bis auf den asymmetrischen Spitzpunkt den sich aus der Klassifikation bezüglich g ergebenden Funktionen. Beim asymmetrischen Spitzpunkt wurde eine Funktion bei Einführung eines zusätzlichen Parameters in 2 Funktionen aufgeteilt.

- Griewank/Reddien [GR89] untersuchten 1989 lokal glatte Kurven singulärer Punkte in Banachräumen, deren Projektion in einen zweidimensionalen Parameterraum einen Spitzpunkt enthält. Dabei sei $\partial_x F(x,\lambda,\alpha)$ ein Operator mit Fredholmindex ≥ 0. Im Endlichdimensionalen bedeutet dies $n \geq m$. Weiter sei $p+k=2$, $q=1$ und $l=0$. Beispiele dafür sind Kurven singulärer Punkte, die einen Hysteresepunkt, d.h. einen zweifachen Rückkehrpunkt, $(m=n)$ bzw. einen Spitzpunkt $(m=n-1)$ enthalten. Für diese singulären Punkte und für TAKENS-BOGDANOV-Punkte $(m=n)$ wurden erweiterte Systeme entwickelt. Anders als in [GR84] wurde hier

  $$B = \begin{bmatrix} \partial_x F & D \\ T & 0 \end{bmatrix} \tag{2.67}$$

  gesetzt. Der dort erwähnte Unterschied bei den Vorzeichen zur Funktion g aus (2.37) tritt jedoch auch hier auf. Unter Beachtung dieses Unterschiedes wurden für Hysteresepunkte die charakterisierenden Funktionen $\partial_\xi g = -N_{n+1,1}^T B^{-1} N_{n+1,1}$ und $\partial_\xi^2 g$, multipliziert mit einer nichtverschwindenden Konstante, verwendet. Für Spitzpunkte entspricht dies $\partial_\xi g$ und $\partial_\xi^2 g$, angewendet auf gewisse Richtungen.

- Dai/Rheinboldt [DR90] zeigten 1990, daß sich das erweiterte System aus [GR84] für einfache Rückkehrpunkte auch zur Berechnung von Mannigfaltigkeiten einfacher Rückkehrpunkte bezüglich spezieller Koordinatenrichtungen verwenden läßt. Dabei wird $q=1$ und $l=0$ vorausgesetzt. Die Raumdimensionen lassen sich als $m=n$ und $p+k \geq 1$ interpretieren, wobei die Zustandsvariablen und die Parameter nicht von vornherein festgelegt zu sein brauchen. Dabei wird eine andere Klassifikation singulärer Punkte als in dieser Arbeit verwendet. Diese Klassifikation und die Ergebnisse einiger geometrischer Untersuchungen wurden von Fink/Rheinboldt [FR87] übernommen. In [FR87] wurde außerdem ein erweitertes System angegeben, deren charakterisierenden Funktionen sich aus den Skalarprodukten eines konstanten Vektors mit den Nullraumvektoren von ∂F ergeben. Diese Skalarprodukte können als die Ableitungen von ϑ bzw. $\boldsymbol{\vartheta}$ interpretiert werden, falls ϑ die entsprechende Linearkombination der Parameter ist.

- 1990 beschrieb Pönisch [Pön90] ein erweitertes System für Verzweigungspunkte $(m=n,\ p=k=q=1,\ l=0,\ \vartheta=\alpha)$ mit den charakterisierenden Funktionen $\partial_{\boldsymbol{\xi}}\boldsymbol{\alpha}$ und $-\partial_\lambda \boldsymbol{\alpha}$, vergleiche Bemerkung 2.47 (iv).

- Im ZAMM-Tagungsband 1994 schlug Pönisch [Pön94] erweiterte Systeme für die singulären Punkte aus 2.3.1 und 2.3.3 mit der Kodimension ≤ 2 vor. δ ist dabei variabel. Als charakterisierende Gleichungen wurden für jeden singulären Punkt jeweils leichte Abwandlungen aller definierenden Gleichungen verwendet. Je nach Punkt wurden als charakterisierende Funktionen $\partial_\xi g$, $-\partial_\xi^2 g$, $-\partial_\lambda g$, $\partial_\xi^3 g$ bzw. $D_2(g)_{(\xi,\lambda)} = \det \partial^2_{(\xi,\lambda)} g$ genutzt.

- Für die gleichen singulären Punkte wie in [Pön94] und zusätzlich für singuläre Punkte aus 2.3.3 mit den Normalformen $\pm\boldsymbol{\xi}^{\ell+1} \pm \lambda\boldsymbol{\xi}$ mit $\ell > 2$ und $\pm\boldsymbol{\xi}^3 \pm \lambda^2$ stellte Schnabel [Sch94] 1994 erweiterte Systeme auf. Dabei beschränkte er sich auf den Fall $k=0$. In den charakterisierenden Funktionen wurden Ableitungen von ϑ verwendet. Bei den Rückkehrpunkten aus 2.3.1 sind dies Ableitungen von λ, beim einfachen Rückkehrpunkt $\partial_\xi \lambda$, beim zweifachen Rückkehrpunkt $-\partial_\xi^2 \lambda$ und beim dreifachen Rückkehrpunkt $-\partial_\xi^3 \lambda$. Die erweiterten Systeme haben die Form (2.8). Dagegen ist bei den

singulären Punkten aus 2.3.3 ϑ gerade die Funktion g mit $\delta = 0$. Die erweiterten Systeme haben die Form (2.10). Als charakterisierende Funktionen wurden bei der Normalform $\pm\boldsymbol{\xi}^3 \pm \lambda^2 \quad - \partial_\xi^2 g$ und $-\partial_\lambda g$, bei der Normalform $\pm\boldsymbol{\xi}^2 \pm \lambda^3 \quad \partial_\xi g$ und $\det(-\partial^2_{(\xi,\lambda)} g)$ und bei den übrigen betrachteten Normalformen $\partial_\xi g$ und $-\partial_\lambda g$, vergleiche Bemerkung 2.47 (iv), verwendet.

- Griewank/Reddien [GR96] verallgemeinerten 1996 das System aus [GR84] für verallgemeinerte Rückkehrpunkte mit $n \geq m$, $p = q = 1$ und $k = l = 0$. Als charakterisierende Funktion wurde wieder $\partial_\xi g$ verwendet. Dabei gilt die bei [GR84] gemachte Bemerkung zu den Vorzeichen auch hier. In dieser Arbeit wurden die beim Newtonverfahren entstehenden linearen Systeme jedoch nicht exakt gelöst.

- 1997 verallgemeinerten Allgower/Schwetlick [AS97] das System von Griewank/Reddien [GR84] für spezielle Verzweigungspunkte bzw. Einsiedlerpunkte auf den Fall $q \leq 2$ und beliebiges δ. Dabei ersetzten sie gemäß der Interpretation in [GR84] (x, λ) durch x. Dies führt zu $m = n - 1$, $p = k = 0$ und $q = l = 1$. Sie zeigten außerdem, daß die Systeme von Pönisch [Pön85] und Janovský [Jan89] für Verzweigungspunkte mit $m = n$ und $q = 1$ Spezialfälle ihres Systems sind. Außerdem wurden in einem Spezialfall die beiden charakterisierenden Funktionen mit unterschiedlichen Werten von δ und zwar einerseits mit $\delta = -\begin{bmatrix} 0 & 1 \end{bmatrix}^T$ und andererseits mit $\delta = -\begin{bmatrix} 1 & 0 \end{bmatrix}^T$ bestimmt.

- Unabhängig von Allgower/Schwetlick [AS97] betrachtete Shen [She97] ebenfalls 1997 das System aus Griewank/Reddien [GR84] für spezielle Verzweigungspunkte bzw. Einsiedlerpunkte. Auch Shen ersetzte (x, λ) durch x, was zu $m = n - 1$, $p = k = 0$ und $q = l = 1$ führt. Im Ausgangssystem ($m = n$, $p = l = 1$, $k = 0$) entspricht dies ebenfalls $q \leq 2$. Shen setzte jedoch $\delta = 0$. Für D nutzte Shen den Linkssingulärvektor zum kleinsten Singulärwert von ∂F in einem gegebenen Startpunkt aus einer Umgebung des gesuchten singulären Punktes. Für die Zeilen von T wählte Shen den Rechtssingulärvektor ebenfalls zum kleinsten Singulärwert und den Nullraumvektor der gleichen Matrix. Als charakterisierende Funktionen verwendete Shen $\partial_\xi g = N_{n+1,1}^T \, B^{-1} \, N_{n+1,2}$. Shen zeigte weiterhin, daß die Matrix B für alle Punkte aus einer hinreichend kleinen Umgebung des singulären Punkt regulär ist, falls der gegebene Startpunkt ebenfalls in dieser Umgebung liegt.

- In [PSS97b], siehe auch [PSS97a] für eine Kurzfassung, wurde 1997 das erweiterte System aus [Sch94] für singuläre Punkte mit der Normalform $\pm\boldsymbol{\xi}^{\ell+1} \pm \lambda\boldsymbol{\xi}$ mit $\ell \geq 2$ vorgeschlagen.

- Govaerts [Gov97a] beschrieb 1997 erweiterte Systeme für die singulären Punkte aus 2.3.1 und 2.3.3. Als charakterisierende Gleichungen verwendete er jeweils alle definierenden Gleichungen mit der Funktion g, wobei er $\delta = 0$ setzte.

- Die erweiterten Systeme aus [Sch94] für Rückkehrpunkte wurden 1999 von Pönisch/Schnabel/Schwetlick [PSS99] für beliebige Rückkehrpunkte aus 2.3.1 mit beliebigen Werten von $k < \ell$ verallgemeinert. Jedoch wurde hier die Funktion $\boldsymbol{\vartheta} = \boldsymbol{\lambda}$ statt der Funktion ϑ verwendet. Als charakterisierende Gleichungen wurden, bis auf das Vorzeichen, $k+1$ definierende Gleichungen ausgewählt, wobei eine charakterisierende

Funktion bei der Normalform $\boldsymbol{\vartheta} = \pm\boldsymbol{\xi}^{\ell+1}$ gerade $\pm\partial_{\boldsymbol{\xi}}^{\ell}\boldsymbol{\lambda}$ ist. Eine Kurzfassung dieser Arbeit ist [PSS98].

Kapitel 3

Ränderung von Matrizen

Wegen den Sätzen 2.29, 2.35 und 2.37 können die Matrizen $\bar{D}$ und $\boldsymbol{T} = \begin{bmatrix} T & T_2 \end{bmatrix}$ beliebig gewählt werden, sofern (2.9) und (2.27) erfüllt sind. Deshalb sollen in diesem Abschnitt einige Untersuchungen zu diesen Ränderungen erfolgen. Shen [She97] verwendete in seinem System bei $m = n - 1$ und $q = l = 1$ die Singulärvektoren zum kleinsten Singulärwert bzw. Nullraumvektoren. Dai/Rheinboldt [DR90, S. 442] schlugen für ihr System bei $m = n$, $q = 1$ u. a. die normierten Singulärvektoren zum kleinsten Singulärwert im aktuellen Punkt vor. Diese Idee soll hier verallgemeinert werden. Allgower/Schwetlick [AS97] zeigten außerdem, daß die Iterationszahl im Newtonverfahren zur Berechnung des singulären Punktes vom Winkel zwischen der Ränderung und diesen Singulärvektoren im singulären Punkt abhängig ist. Diese Singulärvektoren sind im singulären Punkt gerade die Links- und Rechtsnullraumvektoren von $\partial F(x^*) = \partial_x F(x^*)$. Bei großem Winkel zwischen dem von der Ränderung aufgespannten Teilraum und dem von den entsprechenden Singulärvektoren aufgespannten Teilraum kann sich die Anzahl der benötigten Iterationen vergrößern. Zur Verkleinerung des Winkels zwischen den Teilräumen schlugen Allgower/Schwetlick [AS97] die Aufdatierung der Ränderungen vor. Eine geänderte Variante dieser Idee wird in diesem Kapitel ebenfalls vorgestellt. Weiterhin läßt sich, ausgehend von Elsner/He/Mehrmann [EHM95, Abschnitt 4], zeigen, daß durch die Ränderung mit gewissen Vielfachen der Singulärvektoren zu den kleinsten Singulärwerten die Konditionszahl der geränderten Matrix minimal im Vergleich zu allen geränderten Matrizen gleicher Dimension ist. Eine möglichst kleine Konditionszahl ist günstig für eine kleine Fehlerfortpflanzung, vergleiche z. B. Kiełbasiński/Schwetlick [KS88, S. 130]. Es gibt jedoch noch weitere Ränderungen, bei denen die Konditionszahl minimal ist, vergleiche Abschnitt 3.1.

Außer in vielen im Unterabschnitt 2.4.3 erwähnten Systemen werden auch in weiteren Systemen charakterisierende Funktionen verwendet, die sich aus der Lösung geränderter Systeme ergeben. Deshalb ist auch dort eine kleine Konditionszahl der Systemmatrix günstig. Auch in diesen Fällen können die Ergebnisse dieses Kapitels verwendet werden. Dies betrifft z. B. die Berechnung von Takens-Bogdanov-Punkten in [Pön91] und [Pön92] mit $n+q = m+1$ und $l = 0$. In [Jan94] und [JP95] erfolgt die Berechnung von Takens-Bogdanov-Punkten und in [Gov93] von Takens-Bogdanov-Typ-Verzweigungen der Kodimension i mit $n = m$, $q = 1$, der Verwendung von g statt ϑ und $\delta = 0$. Attili verwendete solche charakterisierenden Funktionen in [Att92] im Fall $n = m$, $p = 1$, $q = l = 2$ und $k = 0$. Dabei faßte Attili wie Allgower/Schwetlick in [AS97] x und λ zusammen. Dies kann als $n = m + 1$, $p = k = 0$ und $q = l = 2$ interpretiert werden. Auch Attili setzte $\delta = 0$. In

[SS97a] wurden Periodenverdopplungspunkte mit $m = n$, $q = l = 1$ und $\delta = 0$ berechnet, vergleiche Bemerkung 2.42 (iii). Kunkel [Kun89] verwendete geränderte Matrizen bei der Berechnung singulärer Punkte zur Ermittlung von Pseudoinversen. Dabei sind die in [Pön91] und [Pön92] betrachteten Fälle trivial, da dort mit geeigneten Nullraumvektoren gerändert werden kann, um eine minimale Konditionszahl zu erreichen. Diese lassen sich durch Lösung eines Gleichungssystems bestimmen. In den übrigen Fällen wird $\delta = 0$ gesetzt. Im singulären Punkt selbst ist dies erfüllt, wenn mit Nullraumvektoren gerändert wird. Außerhalb des singulären Punktes ist dagegen $\operatorname{rank} \partial_x F$ bzw. $\operatorname{rank} \partial_{(x,\theta)} \boldsymbol{F}$ i. allg. größer als im singulären Punkt, weshalb $\delta = 0$ eine Einschränkung ist. Dies wird jedoch hier nicht weiter untersucht.

Im folgenden Abschnitt 3.1 werden Ränderungen festgelegter Dimension gegebener Matrizen betrachtet. Dabei werden notwendige und hinreichende Bedingungen für diese Ränderungen angegeben, so daß die Konditionszahl der geränderten Matrix gegenüber allen Ränderungen gleicher Dimension minimal ist. Elsner/He/Mehrmann [EHM95, Abschnitt 4] betrachteten den Fall einer zeilenregulären nichtquadratischen Rechteckmatrix. An diese Matrix fügten sie weitere Zeilen an, so daß die geränderte Matrix quadratisch ist. Sie zeigten notwendige und hinreichende Bedingungen für die Ränderungsmatrix, so daß die Konditionszahl der geränderten Matrix gegenüber allen derartig gewonnenen quadratischen Matrizen minimal ist. Im Abschnitt 3.1 werden nun auch die übrigen Fälle betrachtet. Diese Betrachtungen scheinen noch nicht veröffentlicht zu sein. Wie bereits erwähnt wurde, kann eine solche optimale Ränderung auch mit gewissen Vielfachen der Singulärvektoren zu den kleinsten Singulärwerten erfolgen. Deshalb sollen in den übrigen Abschnitten dieses Kapitels solche Vielfachen ermittelt werden. Im Abschnitt 3.2 werden dazu Varianten der inversen Teilraumiteration für Singulärwerte beschrieben und untersucht. Dabei werden die Ergebnisse aus [SS97b] für die inverse Iteration für Singulärwerte übertragen. Ein Vergleich mit einigen bekannten Verfahren erfolgt zu Beginn jenes Abschnitts. Gegenüber den dort erwähnten Verfahren ist das im Abschnitt 3.2 erwähnte Verfahren neu. Bei der inversen Teilraumiteration konvergieren die Iterierten linear. Deshalb werden im letzten Abschnitt dieses Kapitels einige Überlegungen angestellt, wie das lokal überlinear konvergente Newtonverfahren zur Berechnung der entsprechenden Singulärwerte und -vektoren verwendet werden kann. Diese basieren auf den entsprechenden Überlegungen für Eigenwerte aus [SL97] und [LST98].

3.1 Minimierung der Konditionszahl durch Ränderung

In diesem Abschnitt sollen notwendige und hinreichende Bedingungen für die Ränderungsmatrizen bestimmt werden, so daß die geränderte Matrix minimale Konditionszahl besitzt. Ein Teil dieses Problems wurde von Elsner/He/Mehrmann [EHM95, Abschnitt 4] gelöst. Diese Lösung wird weiter untersucht. Weiterhin wird auch der in [EHM95] nicht betrachtete Teil gelöst.

3.1.1 Notwendige und hinreichende Bedingungen für die Minimalität

Es sei $A := \partial_{(x,\theta)} F \in \mathbb{R}^{m \times (n+q-l)}$ in einem gegebenen Punkt (x, λ, α). Gesucht sind Ränderungen $\bar{D} \in \mathbb{R}^{m \times l}$ und $\boldsymbol{T} \in \mathbb{R}^{(n-m+q) \times (n+q)}$ so, daß $\begin{bmatrix} A & \bar{D} \\ \boldsymbol{T} & \end{bmatrix}$ minimale Konditionszahl besitzt. Dabei bedeutet wieder $l = 0$, daß $\bar{D}$ nicht existiert und somit die geränderte Ma-

trix die Form $\begin{bmatrix} A \\ \boldsymbol{T} \end{bmatrix}$ hat. Der Fall $l = m$ wird zum Schluß behandelt und deshalb vorerst ausgeschlossen.

Definition 3.1. Es seien $\sigma_i(C),\ i = 1, \ldots, \max\{m, n\}$ die Singulärwerte einer Matrix $C \in \mathbb{R}^{m \times n}$ mit

$$\sigma_1(C) \geq \sigma_2(C) \geq \ldots \geq \sigma_{\min\{m,n\}}(C) \geq 0 = \sigma_{\min\{m,n\}+1}(C) = \ldots = \sigma_{\max\{m,n\}}(C)\,. \tag{3.1}$$

Mit den Singulärwertzerlegungen für die Matrizen $A \in \mathbb{R}^{m \times (n+q-l)}$ und $\begin{bmatrix} A & \bar{D} \end{bmatrix} \in \mathbb{R}^{m \times (n+q)}$ sei nun die folgende Definition erfüllt.

Definition 3.2. Es sei $A = U \Sigma V^T$ die Singulärwertzerlegung von A und $\begin{bmatrix} A & \bar{D} \end{bmatrix} = \bar{U} \bar{\Sigma} \bar{V}^T$ die Singulärwertzerlegung von $\begin{bmatrix} A & \bar{D} \end{bmatrix}$. Insbesondere sind die Matrizen $U, \bar{U} \in \mathbb{R}^{m \times m}$, $V \in \mathbb{R}^{(n+q-l) \times (n+q-l)}$ und $\bar{V} \in \mathbb{R}^{(n+q) \times (n+q)}$ orthogonal, Σ hat je nach dem Verhältnis von m und $n + q - l$ die Gestalt $\begin{bmatrix} \operatorname{diag}(\sigma_i, i = 1, \ldots, m) & 0 \end{bmatrix}$ oder $\operatorname{diag}(\sigma_i, i = 1, \ldots, m = n + q - l)$ oder $\begin{bmatrix} \operatorname{diag}(\sigma_i, i = 1, \ldots, n + q - l) & 0 \end{bmatrix}^T$ und $\bar{\Sigma} = \begin{bmatrix} \operatorname{diag}(\bar{\sigma}_i, i = 1, \ldots, m) & 0 \end{bmatrix}$. Dabei ist $\sigma_i = \sigma_i(A)$ und $\bar{\sigma}_i = \sigma_i\left(\begin{bmatrix} A & \bar{D} \end{bmatrix}\right)$. Es seien weiterhin $X := U^T \bar{D}$ und $Y := \bar{V}^T \boldsymbol{T}^T$,

$$\begin{aligned} \Sigma_1 &:= \operatorname{diag}(\sigma_i, \sigma_i = \sigma_1)\,, & \Sigma_2 &:= \operatorname{diag}(\sigma_i, \sigma_1 > \sigma_i > \sigma_{m-l})\,, \\ \Sigma_3 &:= \operatorname{diag}(\sigma_i, \sigma_i = \sigma_{m-l})\,, & \Sigma_4 &:= \operatorname{diag}(\sigma_i, \sigma_i < \sigma_{m-l}, i \leq m) \end{aligned}$$

und X_1, X_2, X_3 bzw. X_4 die entsprechenden Zeilen von X. Analog sei

$$\begin{aligned} \bar{\Sigma}_1 &:= \operatorname{diag}(\bar{\sigma}_i, \bar{\sigma}_i = \bar{\sigma}_1)\,, & \bar{\Sigma}_2 &:= \operatorname{diag}(\bar{\sigma}_i, \bar{\sigma}_1 > \bar{\sigma}_i > \bar{\sigma}_m)\,, \\ \bar{\Sigma}_3 &:= \operatorname{diag}(\bar{\sigma}_i, \bar{\sigma}_i = \bar{\sigma}_m, i \leq m)\,, & \bar{\Sigma}_4 &:= 0 \in \mathbb{R}^{(n-m+q) \times (n-m+q)} \end{aligned}$$

und Y_1, Y_2, Y_3 bzw. Y_4 die entsprechenden Zeilen von Y. Schließlich seien $\hat{\sigma}_i = \sigma_i\left(\begin{bmatrix} A & \bar{D} \\ \boldsymbol{T} & \end{bmatrix}\right)$.

Bemerkung 3.3.

(i) In [EHM95] wurde der Fall $l = 0$ betrachtet. Dort sind die Y_i gerade die Transponierten der hier eingeführten Y_i.

(ii) Falls kein i mit $\sigma_1 > \sigma_i > \sigma_{m-l}$ existiert, ist auch Σ_2 nicht vorhanden. Analoge Aussagen gelten für Σ_4 und für $\bar{\Sigma}_2$.

(iii) Falls sogar $\sigma_1 = \sigma_{m-l}$ gilt, ist nicht nur Σ_2 und damit auch X_2 nicht vorhanden, sondern auch $\Sigma_1 = \Sigma_3$ und $X_1 = X_3$. Analoge Aussagen gelten für $\bar{\sigma}_1 = \bar{\sigma}_m$.

(iv) Die Eigenwerte, vergleiche [GVL89, P 7.1.6, S. 340], und damit auch die Singulärwerte einer Matrix hängen stetig von ihren Elementen ab. Wegen der Glattheit von F, (2.3), (2.48), Bemerkung 2.42 und der Definition von $A = \partial_{(x,\theta)} F = \partial_{(x,\theta)} \boldsymbol{F}$ gilt somit in einer Umgebung des singulären Punktes $\sigma_{m-l+1} < \sigma_{m-l}$. In diesem Fall ist $X_4 \in \mathbb{R}^{l \times l}$ quadratisch, $\Sigma_3 = \operatorname{diag}(\sigma_i, \sigma_i = \sigma_{m-l}, i \leq m - l)$ und $\Sigma_4 = \operatorname{diag}(\sigma_i, m - l < i \leq m)$.

Satz 3.4.
Für $l < m$ und beliebige Matrizen $\bar{D} \in \mathbb{R}^{m\times l}$ und $\boldsymbol{T} \in \mathbb{R}^{(n-m+q)\times(n+q)}$ gilt, falls $\begin{bmatrix} A & \bar{D} \\ \boldsymbol{T} & \end{bmatrix}$ *regulär ist:*

$$\frac{\hat{\sigma}_1}{\hat{\sigma}_{n+q}} = \frac{\sigma_1\left(\begin{bmatrix} A & \bar{D} \\ \boldsymbol{T} & \end{bmatrix}\right)}{\sigma_{n+q}\left(\begin{bmatrix} A & \bar{D} \\ \boldsymbol{T} & \end{bmatrix}\right)} \geq \frac{\bar{\sigma}_1}{\bar{\sigma}_m} = \frac{\sigma_1\left(\begin{bmatrix} A & \bar{D} \end{bmatrix}\right)}{\sigma_m\left(\begin{bmatrix} A & \bar{D} \end{bmatrix}\right)} \geq \frac{\sigma_1}{\sigma_{m-l}} = \frac{\sigma_1(A)}{\sigma_{m-l}(A)} \tag{3.2}$$

und die kleinste mit $\bar{D}$ und $\boldsymbol{T}$ zu erreichende Konditionszahl beträgt

$$\frac{\hat{\sigma}_1}{\hat{\sigma}_{n+q}} = \frac{\sigma_1\left(\begin{bmatrix} A & \bar{D} \\ \boldsymbol{T} & \end{bmatrix}\right)}{\sigma_{n+q}\left(\begin{bmatrix} A & \bar{D} \\ \boldsymbol{T} & \end{bmatrix}\right)} = \frac{\bar{\sigma}_1}{\bar{\sigma}_m} = \frac{\sigma_1\left(\begin{bmatrix} A & \bar{D} \end{bmatrix}\right)}{\sigma_m\left(\begin{bmatrix} A & \bar{D} \end{bmatrix}\right)} = \frac{\sigma_1}{\sigma_{m-l}} = \frac{\sigma_1(A)}{\sigma_{m-l}(A)}\,. \tag{3.3}$$

Dies wird genau dann erreicht, wenn

$$\begin{aligned} \sigma_1\left(\begin{bmatrix} A & \bar{D} \\ \boldsymbol{T} & \end{bmatrix}\right) &= \sigma_1\left(\begin{bmatrix} A & \bar{D} \end{bmatrix}\right) = \sigma_1(A) && \text{und} \\ \sigma_{n+q}\left(\begin{bmatrix} A & \bar{D} \\ \boldsymbol{T} & \end{bmatrix}\right) &= \sigma_m\left(\begin{bmatrix} A & \bar{D} \end{bmatrix}\right) = \sigma_{m-l}(A)\,. \end{aligned} \tag{3.4}$$

Dies ist genau dann erfüllt, wenn

$$\begin{aligned} &\bar{\sigma}_1^2 I_{n+q} - \begin{bmatrix} A & \bar{D} \\ \boldsymbol{T} & \end{bmatrix}\begin{bmatrix} A & \bar{D} \\ \boldsymbol{T} & \end{bmatrix}^T, && \sigma_1^2 I_m - \begin{bmatrix} A & \bar{D} \end{bmatrix}\begin{bmatrix} A & \bar{D} \end{bmatrix}^T && \text{und} \\ &\begin{bmatrix} A & \bar{D} \\ \boldsymbol{T} & \end{bmatrix}\begin{bmatrix} A & \bar{D} \\ \boldsymbol{T} & \end{bmatrix}^T - \bar{\sigma}_m^2 I_{n+q}\,, && \begin{bmatrix} A & \bar{D} \end{bmatrix}\begin{bmatrix} A & \bar{D} \end{bmatrix}^T - \sigma_{m-l}^2 I_m \end{aligned} \tag{3.5}$$

positiv semidefinit sind.

Beweis. (3.2) folgt aus der entsprechenden Beziehung für $l = 1$ und $n-m+q = 1$. Diese folgt aus der Interlacing Property für Singulärwerte, siehe z. B. Björck [Bjö96, Theorem 1.2.9]. Diese Beziehung läßt sich aus der entsprechenden Beziehung für Eigenwerte, siehe z. B. Golub/Van Loan [GVL89, Corollary 8.1.4, S. 411], herleiten. Wegen der erwähnten Interlacing Property für Singulärwerte ist σ_1 der kleinstmögliche Wert von $\hat{\sigma}_1$ und σ_{m-l} der größtmögliche Wert von $\hat{\sigma}_{n+q}$. Bei

$$\begin{aligned} X^T &= \left[0 \quad \operatorname{diag}\left(\sqrt{\sigma_1^2 - \sigma_{m-l+i}^2}, i = 1, \ldots, l\right)\right] && \text{und} \\ Y^T &= \left[0 \quad \operatorname{diag}(\sigma_1)\right] = \sigma_1 N_{n+q,n-m+q}^T \end{aligned} \tag{3.6}$$

ist:

$$\begin{aligned} &\bar{\sigma}_i = \sigma_1,\ i = 1, \ldots, l, && \bar{\sigma}_{i+l} = \sigma_i,\ i = 1, \ldots, m-l, \\ &\hat{\sigma}_i = \sigma_1,\ i = 1, \ldots, n-m+q+l \quad \text{und} && \hat{\sigma}_{i+n-m+q+l} = \sigma_i,\ i = 1, \ldots, m-l. \end{aligned}$$

Somit ist (3.4) stets erreichbar und optimal. Bei $\hat{\sigma}_{n+q} = \bar{\sigma}_m = \sigma_{m-l} > 0$ ist dann auch (3.3) erfüllt und optimal. Wegen der Optimalität von (3.4) ist (3.4) auch notwendig für (3.3).

(3.4) läßt sich auch unter Beachtung der Ordnung (3.1) der Singulärwerte und der Optimalität von (3.4) auch als

$$\forall i = 1, \ldots, n+q : \bar{\sigma}_1 \geq \hat{\sigma}_i \geq \bar{\sigma}_m \quad \text{und} \quad \forall i = 1, \ldots, m : \sigma_1 \geq \bar{\sigma}_i \geq \sigma_{m-l}$$

schreiben. Dies bedeutet gerade, daß

(3.7)
$$\bar{\sigma}_1^2 I_{n+q} - \operatorname{diag}\left(\hat{\sigma}_i^2, i = 1, \ldots, n+q\right), \qquad \sigma_1^2 I_m - \operatorname{diag}\left(\bar{\sigma}_i^2, i = 1, \ldots, m\right) \qquad \text{und}$$
$$\operatorname{diag}\left(\hat{\sigma}_i^2, i = 1, \ldots, n+q\right) - \bar{\sigma}_m^2 I_{n+q}, \qquad \operatorname{diag}\left(\bar{\sigma}_i^2, i = 1, \ldots, m\right) - \sigma_{m-l}^2 I_m$$

positiv semidefinit sind. Durch Multiplikation der rechten Spalte von (3.7) von links mit $\bar{U}$ und von rechts mit $\bar{U}^T$ ergibt sich die rechte Spalte von (3.5). Durch eine analoge Transformation der linken Spalte von (3.7) ergibt sich die linke Spalte von (3.5). Daraus folgt die Aussage des Satzes. □

Bemerkung 3.5.

(i) Im Fall $l = 0$ ist $\begin{bmatrix} A & \bar{D} \end{bmatrix} = A$ und somit $\bar{\sigma}_i = \sigma_i$, $i = 1, \ldots, m$. (3.2) wird dann zu $\frac{\hat{\sigma}_1}{\hat{\sigma}_{n+q}} \geq \frac{\sigma_1}{\sigma_m}$, (3.3) zu $\frac{\hat{\sigma}_1}{\hat{\sigma}_{n+q}} = \frac{\sigma_1}{\sigma_m}$, (3.4) zu $\hat{\sigma}_1 = \sigma_1$ und $\hat{\sigma}_{n+q} = \sigma_m$ und (3.5) zu

$$\sigma_1^2 I_{n+q} - \begin{bmatrix} A \\ \boldsymbol{T} \end{bmatrix} \begin{bmatrix} A \\ \boldsymbol{T} \end{bmatrix}^T \quad \text{und} \quad \begin{bmatrix} A \\ \boldsymbol{T} \end{bmatrix} \begin{bmatrix} A \\ \boldsymbol{T} \end{bmatrix}^T - \sigma_m^2 I_{n+q}\,.$$

(ii) Im Fall $l = m$ ist wegen Definition 2.1 und (2.3) $q = m$. Weiterhin wird (3.2) zu $\frac{\hat{\sigma}_1}{\hat{\sigma}_{n+m}} \geq \frac{\bar{\sigma}_1}{\bar{\sigma}_m} \geq 1$, (3.3) zu $\frac{\hat{\sigma}_1}{\hat{\sigma}_{n+m}} = \frac{\bar{\sigma}_1}{\bar{\sigma}_m} = 1$ und (3.4) zu

(3.8) $$\hat{\sigma}_1 = \bar{\sigma}_1 = \bar{\sigma}_m = \hat{\sigma}_{n+m}\,,$$

während die positive Semidefinitheit von (3.5) zu

(3.9) $$\bar{\sigma}_1^2 I_{n+m} = \begin{bmatrix} A & \bar{D} \\ \boldsymbol{T} & \end{bmatrix} \begin{bmatrix} A & \bar{D} \\ \boldsymbol{T} & \end{bmatrix}^T \quad \text{und} \quad \bar{\sigma}_1^2 I_m = \begin{bmatrix} A & \bar{D} \end{bmatrix} \begin{bmatrix} A & \bar{D} \end{bmatrix}^T$$

zusammengefaßt wird. Der Beweis erfolgt analog zum Beweis von Satz 3.4, wobei $\bar{\sigma}_1 \geq \bar{\sigma}_m$ aus (3.1) folgt. Speziell können in (3.6) die Matrizen $X = \operatorname{diag}\left(\sqrt{\bar{\sigma}_1^2 - \sigma_i^2}, i = 1, \ldots, m\right)$ und $Y = \bar{\sigma}_1 N_{n+m,n}$ mit $\bar{\sigma}_1 \geq \sigma_1$ und $\bar{\sigma}_1 > 0$ gewählt werden, woraus $\bar{\sigma}_1 = \bar{\sigma}_i > 0, i = 1, \ldots, m$ und $\bar{\sigma}_1 = \hat{\sigma}_i, i = 1, \ldots, n+m$ folgt.

3.1.2 Bedingungen unter Nutzung von Pseudoinversen

Die notwendigen und hinreichenden Bedingungen (3.5) sollen nun analog zu [EHM95, Proposition 4.1] umformuliert werden. Dabei werde im Folgenden in diesem Unterabschnitt $\sigma_{m-l+1} < \sigma_{m-l}$ gefordert. Wegen Bemerkung 3.3 (iv) ist dies in einer Umgebung des singulären Punktes keine Einschränkung.

Das folgende Lemma 3.6 wird im Beweis des folgenden Satzes 3.7 genutzt.

Lemma 3.6.
Es sei $C \in \mathbb{R}^{m\times n}$ eine beliebige Matrix. Dann gilt:

$$\forall z \in \mathbb{R}^m : z^T z - z^T C\, C^T z \geq 0 \quad \Longleftrightarrow \quad \begin{array}{l} I_n - C^T C \text{ ist positiv semidefinit, d. h.} \\ \forall z \in \mathbb{R}^n : z^T z \geq z^T C^T C\, z\,. \end{array}$$

Beweis. Wegen $\left\|C^T\right\|_2 = \|C\|_2$, vergleiche z. B. Kiełbasiński/Schwetlick [KS88, Ü 1.2.9, (33), S. 44], gilt

$$\begin{aligned} \forall z \in \mathbb{R}^m : z^T z - z^T C\, C^T z \geq 0 \;&\Leftrightarrow\; 1 \geq \max_{z\neq 0} \frac{z^T C\, C^T z}{z^T z} = \left\|C^T\right\|_2^2 = \|C\|_2^2 = \max_{z\neq 0} \frac{z^T C^T C\, z}{z^T z} \\ &\Leftrightarrow\; \forall z \in \mathbb{R}^n : z^T z \geq z^T C^T C\, z\,, \quad \text{d. h.,} \end{aligned}$$

$I - C^T C$ ist positiv semidefinit. □

Satz 3.7.
Bei $1 \leq l < m$ und $\sigma_{m-l+1} < \sigma_{m-l}$ ist genau dann $\bar\sigma_1 = \sigma_1$ und $\bar\sigma_m = \sigma_{m-l}$, wenn $X_1 = 0$, $X_3 = 0$, X_4 regulär, $I_l - \bar D^T \left(\sigma_1^2 I_m - A\, A^T\right)^+ \bar D$ positiv semidefinit und $I_l + \bar D^T \left(A\, A^T - \sigma_{m-l}^2 I_m\right)^+ \bar D$ negativ semidefinit ist.

Bemerkung 3.8.

(i) Für $l = 0$ ist wegen Bemerkung 3.5 $\bar\sigma_1 = \sigma_1$ und $\bar\sigma_m = \sigma_{m-l}$ trivial.

(ii) Offensichtlich sind $\sigma_1^2 I_m - \Sigma\, \Sigma^T = \operatorname{diag}\left(\sigma_1^2 - \sigma_i^2\right)$, $\Sigma_2\, \Sigma_2^T - \sigma_{m-l}^2 I$ und $\sigma_{m-l}^2 I_l - \Sigma_4\, \Sigma_4^T$ positiv semidefinite bzw. positiv definite Diagonalmatrizen. Die Pseudoinverse, siehe z. B. [KS88, S. 249] für die Definition, bei semidefiniten bzw. die Inverse bei definiten Matrizen ist ebenfalls eine Diagonalmatrix, deren nichtverschwindenden Diagonalelemente gerade die Inversen der nichtverschwindenden Diagonalelemente der Ausgangsmatrix sind.

(iii) Wie üblich, vergleiche z. B. [GVL89, P 11.2.4, S. 554], wird in dieser Arbeit unter der Wurzel einer positiv (semi)definiten Matrix gerade die positiv (semi)definite Matrix verstanden, deren Quadrat die Ausgangsmatrix ist. Bei den in Bemerkung 3.8 (ii) eingeführten Matrizen sind die Wurzeln Diagonalmatrizen, deren Diagonalelemente die Wurzeln der Ausgangsdiagonalelemente sind.

Beweis. $\bar\sigma_1 = \sigma_1$ und $\bar\sigma_m = \sigma_{m-l}$ ist wegen Satz 3.4 äquivalent zur positiven Semidefinitheit der auf der rechten Seite von (3.5) stehenden Ausdrücke. Durch Multiplikation dieser Ausdrücke von links mit U^T und von rechts mit U erhält man

$$\sigma_1^2 I_m - \Sigma\, \Sigma^T - X\, X^T \quad \text{und} \quad \Sigma\, \Sigma^T - \sigma_{m-l}^2 I_m + X\, X^T.$$

Aus der positiven Semidefinitheit von $\sigma_1^2 I_m - \Sigma\, \Sigma^T - X\, X^T$ folgt insbesondere für alle z_1

$$0 \leq z_1^T \left(\sigma_1^2 I - \Sigma_1\, \Sigma_1^T - X_1\, X_1^T\right) z_1 = -z_1^T X_1\, X_1^T z_1\,,$$

woraus $X_1 = 0$ folgt.

Wegen der positiven Semidefinitheit von $\Sigma\, \Sigma^T - \sigma_{m-l}^2 I_m + X\, X^T$ und der negativen Definitheit von $\Sigma_4\, \Sigma_4^T - \sigma_{m-l}^2 I_l$ gilt für alle $z_4 \in \ker X_4^T$

$$0 \leq z_4^T \left(\Sigma_4\, \Sigma_4^T - \sigma_{m-l}^2 I_l + X_4\, X_4^T\right) z_4 = z_4^T \left(\Sigma_4\, \Sigma_4^T - \sigma_{m-l}^2 I_l\right) z_4\,,$$

woraus $z_4 = 0$ folgt. Somit ist X_4^T und damit auch X_4 regulär. Daraus folgt für alle z_3 und $z_4 = -X_4^{-T} X_3^T z_3$ wegen

$$\begin{aligned} 0 &\le z_3^T \left(\Sigma_3 \Sigma_3^T - \sigma_{m-l}^2 I\right) z_3 + z_4^T \left(\Sigma_4 \Sigma_4^T - \sigma_{m-l}^2 I_l\right) z_4 + \left(z_3^T X_3 + z_4^T X_4\right) \left(X_3^T z_3 + X_4^T z_4\right) \\ &= z_3^T X_3 X_4^{-1} \left(\Sigma_4 \Sigma_4^T - \sigma_{m-l}^2 I_l\right) X_4^{-T} X_3^T z_3 \end{aligned}$$

$X_4^{-T} X_3^T z_3 = 0$, somit $X_3^T z_3 = 0$ und deshalb $X_3 = 0$.

Nach Definition bedeutet die positive Semidefinitheit von $\sigma_1^2 I_m - \Sigma \Sigma^T - X X^T$

$$z^T \left(\sigma_1^2 I_m - \Sigma \Sigma^T - X X^T\right) z \ge 0 \tag{3.10}$$

für alle $z \in \mathbb{R}^m$. Wegen $X_1 = 0$ hängt der Wert der linken Seite von (3.10) nicht von z_1, der zu Σ_1 und X_1 gehörenden Komponenten von z ab. Deshalb kann o. B. d. A. $z_1 = 0$ gesetzt werden. Deshalb kann z durch $\left[\left(\sigma_1^2 I_m - \Sigma \Sigma^T\right)^{\frac{1}{2}}\right]^+ z$ ersetzt werden, d. h. $z := \left(\sigma_1^2 I_m - \Sigma \Sigma^T\right)^{\frac{1}{2}} z$. Dann ergibt sich aus (3.10) für alle $z \in \mathbb{R}^m$

$$z^T z - z^T \left[\left(\sigma_1^2 I_m - \Sigma \Sigma^T\right)^{\frac{1}{2}}\right]^+ X X^T \left[\left(\sigma_1^2 I_m - \Sigma \Sigma^T\right)^{\frac{1}{2}}\right]^+ z \ge 0 \,.$$

Mit $C := \left[\left(\sigma_1^2 I_m - \Sigma \Sigma^T\right)^{\frac{1}{2}}\right]^+ X$ ergibt sich aus Lemma 3.6, daß (3.10) unter Beachtung von $X_1 = 0$ genau dann erfüllt ist, wenn $I_l - X^T \left(\sigma_1^2 I_m - \Sigma \Sigma^T\right)^+ X$ und damit auch $I_l - \bar{D}^T \left(\sigma_1^2 I_m - A A^T\right)^+ \bar{D}$ positiv semidefinit ist.

Die positive Semidefinitheit von $\Sigma \Sigma^T - \sigma_{m-l}^2 I_m + X X^T$ bedeutet

$$z^T \left(\Sigma \Sigma^T - \sigma_{m-l}^2 I_m + X X^T\right) z \ge 0 \,. \tag{3.11}$$

Im Fall, daß Σ_2 und damit X_2 nicht existieren, ist dafür wegen $X_1 = 0$ und $X_3 = 0$ notwendig und hinreichend, daß

$$z_4^T \left(\Sigma_4 \Sigma_4^T - \sigma_{m-l}^2 I_l\right) z_4 + z_4^T X_4 X_4^T z_4 \ge 0 \tag{3.12}$$

erfüllt ist. Wegen der Regularität von X_4 kann z_4 durch $X_4^{-T} z_4$ ersetzt werden, was der regulären Koordinatentransformation $z_4 := X_4^T z_4$ entspricht. Damit wird aus (3.12)

$$-z_4^T X_4^{-1} \left(\sigma_{m-l}^2 I_l - \Sigma_4 \Sigma_4^T\right) X_4^{-T} z_4 + z_4^T z_4 \ge 0 \,.$$

Mit $C := X_4^{-1} \left(\sigma_{m-l}^2 I_l - \Sigma_4 \Sigma_4^T\right)^{\frac{1}{2}}$ ergibt sich aus Lemma 3.6, daß (3.12) genau dann für alle z_4 erfüllt ist, falls für alle $z \in \mathbb{R}^l$

$$z^T z \ge z^T \left(\sigma_{m-l}^2 I_l - \Sigma_4 \Sigma_4^T\right)^{\frac{1}{2}} X_4^{-T} X_4^{-1} \left(\sigma_{m-l}^2 I_l - \Sigma_4 \Sigma_4^T\right)^{\frac{1}{2}} z$$

gilt. Wenn nun z durch $\left(\sigma_{m-l}^2 I_l - \Sigma_4 \Sigma_4^T\right)^{-\frac{1}{2}} X_4 z$ ersetzt wird, was der regulären Koordinatentransformation $z := X_4^{-1} \left(\sigma_{m-l}^2 I_l - \Sigma_4 \Sigma_4^T\right)^{\frac{1}{2}} z$ entspricht, ergibt sich

$$z^T X_4^T \left(\sigma_{m-l}^2 I_l - \Sigma_4 \Sigma_4^T\right)^{-1} X_4 z \ge z^T z \,.$$

Dies bedeutet, daß in diesem Fall

$$I_l + X_4^T \left(\Sigma_4 \Sigma_4^T - \sigma_{m-l}^2 I_l\right)^{-1} X_4 = I_l + X^T \left(\Sigma \Sigma^T - \sigma_{m-l}^2 I_m\right)^+ X \tag{3.13}$$

negativ semidefinit ist.

Falls dagegen Σ_2 und damit auch X_2 existieren, ist für (3.11) ebenfalls wegen $X_1 = 0$ und $X_3 = 0$ notwendig und hinreichend, daß

$$0 \leq z_2^T \left(\Sigma_2 \Sigma_2^T - \sigma_{m-l}^2 I\right) z_2 + z_4^T \left(\Sigma_4 \Sigma_4^T - \sigma_{m-l}^2 I_l\right) z_4 + \left(z_2^T X_2 + z_4^T X_4\right) \left(X_2^T z_2 + X_4^T z_4\right) \tag{3.14}$$

erfüllt ist. Wenn

$$\begin{pmatrix} z_2 \\ z_4 \end{pmatrix} \quad \text{durch} \quad \begin{bmatrix} \left(\Sigma_2 \Sigma_2^T - \sigma_{m-l}^2 I\right)^{-\frac{1}{2}} & 0 \\ -X_4^{-T} X_2^T \left(\Sigma_2 \Sigma_2^T - \sigma_{m-l}^2 I\right)^{-\frac{1}{2}} & X_4^{-T} \end{bmatrix} \begin{pmatrix} z_2 \\ z_4 \end{pmatrix}$$

ersetzt wird, was der regulären Koordinatentransformation

$$\begin{pmatrix} z_2 \\ z_4 \end{pmatrix} := \begin{bmatrix} \left(\Sigma_2 \Sigma_2^T - \sigma_{m-l}^2 I\right)^{\frac{1}{2}} & 0 \\ X_2^T & X_4^T \end{bmatrix} \begin{pmatrix} z_2 \\ z_4 \end{pmatrix}$$

entspricht, ergibt sich

$$\begin{aligned} 0 \leq & z_2^T z_2 + \begin{pmatrix} z_2^T & z_4^T \end{pmatrix} \begin{bmatrix} -\left(\Sigma_2 \Sigma_2^T - \sigma_{m-l}^2 I\right)^{-\frac{1}{2}} X_2 \\ I_l \end{bmatrix} X_4^{-1} \left(\Sigma_4 \Sigma_4^T - \sigma_{m-l}^2 I_l\right) X_4^{-T} \\ & \begin{bmatrix} -X_2^T \left(\Sigma_2 \Sigma_2^T - \sigma_{m-l}^2 I\right)^{-\frac{1}{2}} & I_l \end{bmatrix} \begin{pmatrix} z_2 \\ z_4 \end{pmatrix} + z_4^T z_4 \,. \end{aligned}$$

Mit $C := \begin{bmatrix} -\left(\Sigma_2 \Sigma_2^T - \sigma_{m-l}^2 I\right)^{-\frac{1}{2}} X_2 \\ I_l \end{bmatrix} X_4^{-1} \left(\sigma_{m-l}^2 I_l - \Sigma_4 \Sigma_4^T\right)^{\frac{1}{2}}$ folgt dann aus Lemma 3.6, daß (3.14) genau dann für alle $\begin{pmatrix} z_2^T & z_4^T \end{pmatrix}$ erfüllt ist, falls für alle $z \in \mathbb{R}^l$

$$\begin{aligned} & z^T z \\ & \geq z^T \left(\sigma_{m-l}^2 I_l - \Sigma_4 \Sigma_4^T\right)^{\frac{1}{2}} X_4^{-T} \left(X_2^T \left(\Sigma_2 \Sigma_2^T - \sigma_{m-l}^2 I\right)^{-1} X_2 + I_l\right) X_4^{-1} \left(\sigma_{m-l}^2 I_l - \Sigma_4 \Sigma_4^T\right)^{\frac{1}{2}} z \end{aligned}$$

gilt. Wenn nun z durch $\left(\sigma_{m-l}^2 I_l - \Sigma_4 \Sigma_4^T\right)^{-\frac{1}{2}} X_4 z$ ersetzt wird, was der regulären Koordinatentransformation $z := X_4^{-1} \left(\sigma_{m-l}^2 I_l - \Sigma_4 \Sigma_4^T\right)^{\frac{1}{2}} z$ entspricht, ergibt sich

$$z^T X_4^T \left(\sigma_{m-l}^2 I_l - \Sigma_4 \Sigma_4^T\right)^{-1} X_4 z \geq z^T \left(X_2^T \left(\Sigma_2 \Sigma_2^T - \sigma_{m-l}^2 I\right)^{-1} X_2 + I_l\right) z \,.$$

Dies bedeutet, daß auch in diesem Fall

$$\begin{aligned} I_l + X_2^T \left(\Sigma_2 \Sigma_2^T - \sigma_{m-l}^2 I\right)^{-1} X_2 + X_4^T \left(\Sigma_4 \Sigma_4^T - \sigma_{m-l}^2 I_l\right)^{-1} X_4 & \\ = I_l + X^T \left(\Sigma \Sigma^T - \sigma_{m-l}^2 I_m\right)^+ X & \end{aligned}$$

negativ semidefinit ist. Zusammen mit (3.13) ergibt sich daraus, daß (3.11) unter Beachtung von $X_3 = 0$ und der Regularität von X_4 genau dann erfüllt ist, wenn $I_l + X^T \left(\Sigma \Sigma^T - \sigma_{m-l}^2 I_m\right)^+ X$ und damit auch $I_l + \bar{D}^T \left(A A^T - \sigma_{m-l}^2 I_m\right)^+ \bar{D}$ negativ semidefinit ist. □

Bemerkung 3.9.

(i) Falls Σ_2 nicht existiert, d.h., falls es kein i mit $\sigma_1 > \sigma_i > \sigma_{m-l}$ gibt, sind unter Beachtung von $X_1 = 0$ und $X_3 = 0$ die Matrizen $I_l - X_4^T \left(\sigma_1^2 I_l - \Sigma_4 \Sigma_4^T\right)^{-1} X_4$ und $X_4^T \left(\sigma_{m-l}^2 I_l - \Sigma_4 \Sigma_4^T\right)^{-1} X_4 - I_l$ positiv semidefinit. (3.10) und (3.11) bedeuten dann, daß $\sigma_1^2 I_l - \Sigma_4 \Sigma_4^T - X_4 X_4^T$ und $\Sigma_4 \Sigma_4^T - \sigma_{m-l}^2 I_l + X_4 X_4^T$ positiv semidefinit sind.

(ii) Falls sogar $\sigma_1 = \sigma_{m-l}$ ist, bedeutet dies $X_4 X_4^T = \sigma_1^2 I_l - \Sigma_4 \Sigma_4^T$.

(iii) Offensichtlich ist I_l positiv definit und $X_2^T \left(\Sigma_2 \Sigma_2^T - \sigma_{m-l}^2 I\right)^{-1} X_2$ positiv semidefinit, weshalb $I_l + X_2^T \left(\Sigma_2 \Sigma_2^T - \sigma_{m-l}^2 I\right)^{-1} X_2$ positiv definit ist. Deshalb gilt für alle $z \in \ker X_4$

$$\begin{aligned} 0 &\geq z^T \left(I_l + X_2^T \left(\Sigma_2 \Sigma_2^T - \sigma_{m-l}^2 I\right)^{-1} X_2\right) z + z^T X_4^T \left(\Sigma_4 \Sigma_4^T - \sigma_{m-l}^2 I_l\right)^{-1} X_4 z \\ &= z^T \left(I_l + X_2^T \left(\Sigma_2 \Sigma_2^T - \sigma_{m-l}^2 I\right)^{-1} X_2\right) z \,, \end{aligned}$$

woraus $z = 0$ folgt. Somit folgt die Regularität von X_4 auch aus der negativen Semidefinitheit von $I_l + \bar{D}^T \left(A A^T - \sigma_{m-l}^2 I_m\right)^+ \bar{D}$.

(iv) Dagegen folgt $X_1 = 0$ und $X_3 = 0$ nicht aus der positiven Semidefinitheit von $I_l - \bar{D}^T \left(\sigma_1^2 I_m - A A^T\right)^+ \bar{D}$ bzw. der negativen Semidefinitheit von $I_l + \bar{D}^T \left(A A^T - \sigma_{m-l}^2 I_m\right)^+ \bar{D}$. Es ist nämlich $X_1^T \left(\sigma_1^2 I - \Sigma_1 \Sigma_1^T\right)^+ X_1 = 0$ und $X_3^T \left(\Sigma_3 \Sigma_3^T - \sigma_{m-l}^2 I\right)^+ X_3 = 0$ unabhängig von X_1 und X_3.

Analog zu Satz 3.7 ergibt sich

Satz 3.10.
Bei $\bar{\sigma}_m > 0$ ist genau dann $\hat{\sigma}_1 = \bar{\sigma}_1$ und $\hat{\sigma}_{n+q} = \bar{\sigma}_m$, wenn $Y_1 = 0$, $Y_3 = 0$,

$$\begin{aligned} &I_{n-m+q} - \boldsymbol{T} \left(\bar{\sigma}_1^2 I_{n+q} - \begin{bmatrix} A & \bar{D} \end{bmatrix}^T \begin{bmatrix} A & \bar{D} \end{bmatrix}\right)^+ \boldsymbol{T}^T \quad \textit{positiv semidefinit} \qquad \textit{und} \\ &I_{n-m+q} + \boldsymbol{T} \left(\begin{bmatrix} A & \bar{D} \end{bmatrix}^T \begin{bmatrix} A & \bar{D} \end{bmatrix} - \bar{\sigma}_m^2 I_{n+q}\right)^+ \boldsymbol{T}^T \quad \textit{negativ semidefinit} \end{aligned}$$

ist. Weiterhin folgt die Regularität von Y_4 sowohl aus $\hat{\sigma}_{n+q} = \bar{\sigma}_m$ als auch aus der negativen Semidefinitheit von $I_{n-m+q} + \boldsymbol{T} \left(\begin{bmatrix} A & \bar{D} \end{bmatrix}^T \begin{bmatrix} A & \bar{D} \end{bmatrix} - \bar{\sigma}_m^2 I_{n+q}\right)^+ \boldsymbol{T}^T$.

Beweis. Wegen Definition 2.1 ist $n-m+q \geq 1$. Bei $\bar{\sigma}_m > 0$ gilt auch $\bar{\sigma}_m > \bar{\sigma}_{m+1} = 0$. Da im Beweis von Satz 3.7 nicht ausgenutzt wurde, daß $n+q > m$ ist, ergibt sich Satz 3.10 analog zu Satz 3.7, wobei Σ durch $\bar{\Sigma}^T$, l durch $n-m+q$, m durch $n+q$, σ_i durch $\bar{\sigma}_i$, $\bar{\sigma}_i$ durch $\hat{\sigma}_i$, X durch Y, $\bar{D}$ durch $\boldsymbol{T}^T$ und A^T durch $\begin{bmatrix} A & \bar{D} \end{bmatrix}$ ersetzt wird. Allerdings ist dabei zu beachten, daß in (3.5) $\begin{bmatrix} A & \bar{D} \\ \boldsymbol{T} & \end{bmatrix} \begin{bmatrix} A & \bar{D} \\ \boldsymbol{T} & \end{bmatrix}^T$ steht, während bei der direkten Übertragung $\begin{bmatrix} A & \bar{D} \\ \boldsymbol{T} & \end{bmatrix}^T \begin{bmatrix} A & \bar{D} \\ \boldsymbol{T} & \end{bmatrix}$ stehen würde. Dies beeinflußt die Eigenschaften jedoch nicht. Die Aussagen zur Regularität von Y_4 ergeben sich analog zu dem Beweis von Satz 3.7 und zur Bemerkung 3.9 (iii). □

Bemerkung 3.11.

(i) Elsner/He/Mehrmann [EHM95] betrachteten komplexe Zahlen. Außerdem forderten sie statt der positiven Semidefinitheit der Matrizen

$$I_{n-m+q} - \boldsymbol{T}\left(\bar{\sigma}_1^2 I_{n+q} - \begin{bmatrix} A & \bar{D}\end{bmatrix}^T \begin{bmatrix} A & \bar{D}\end{bmatrix}\right)^+ \boldsymbol{T}^T \qquad \text{und}$$
$$\boldsymbol{T}\left(\bar{\sigma}_m^2 I_{n+q} - \begin{bmatrix} A & \bar{D}\end{bmatrix}^T \begin{bmatrix} A & \bar{D}\end{bmatrix}\right)^+ \boldsymbol{T}^T - I_{n-m+q}$$

andere Bedingungen, die aber für reelle Zahlen gleichwertig sind. In der in der vorliegenden Arbeit verwendeten Notation forderten sie die Riccati-Ungleichungen

$$\boldsymbol{T}\left(I - \begin{bmatrix} A & \bar{D}\end{bmatrix}^T \left(\begin{bmatrix} A & \bar{D}\end{bmatrix} \begin{bmatrix} A & \bar{D}\end{bmatrix}^T - \bar{\sigma}_1^2 I\right)^+ \begin{bmatrix} A & \bar{D}\end{bmatrix}\right) \boldsymbol{T}^T - \bar{\sigma}_1^2 I \tag{3.15}$$

negativ semidefinit und

$$\boldsymbol{T}\left(I - \begin{bmatrix} A & \bar{D}\end{bmatrix}^T \left(\begin{bmatrix} A & \bar{D}\end{bmatrix} \begin{bmatrix} A & \bar{D}\end{bmatrix}^T - \bar{\sigma}_m^2 I\right)^+ \begin{bmatrix} A & \bar{D}\end{bmatrix}\right) \boldsymbol{T}^T - \bar{\sigma}_m^2 I \tag{3.16}$$

positiv semidefinit. Dies vereinfachten sie unter Beachtung von $Y_1 = 0$ und $Y_3 = 0$ zu $Y_4^T Y_4 - \bar{\sigma}_m^2 \left(I + Y_2^T \left(\bar{\Sigma}_2^2 - \bar{\sigma}_m^2 I\right)^{-1} Y_2\right)$ und $\bar{\sigma}_1^2 \left(I + Y_2^T \left(\bar{\Sigma}_2^2 - \bar{\sigma}_1^2 I\right)^{-1} Y_2\right) - Y_4^T Y_4$ positiv semidefinit.

(ii) (3.15) ergibt ausmultipliziert

$$\boldsymbol{T}\,\boldsymbol{T}^T - \bar{\sigma}_1^2 I - \boldsymbol{T} \begin{bmatrix} A & \bar{D}\end{bmatrix}^T \left(\begin{bmatrix} A & \bar{D}\end{bmatrix} \begin{bmatrix} A & \bar{D}\end{bmatrix}^T - \bar{\sigma}_1^2 I\right)^+ \begin{bmatrix} A & \bar{D}\end{bmatrix} \boldsymbol{T}^T,$$

während (3.16)

$$\boldsymbol{T}\,\boldsymbol{T}^T - \bar{\sigma}_m^2 I - \boldsymbol{T} \begin{bmatrix} A & \bar{D}\end{bmatrix}^T \left(\begin{bmatrix} A & \bar{D}\end{bmatrix} \begin{bmatrix} A & \bar{D}\end{bmatrix}^T - \bar{\sigma}_m^2 I\right)^+ \begin{bmatrix} A & \bar{D}\end{bmatrix} \boldsymbol{T}^T$$

entspricht. Da $\left(\begin{bmatrix} A & \bar{D}\end{bmatrix} \begin{bmatrix} A & \bar{D}\end{bmatrix}^T - \bar{\sigma}_1^2 I\right)^+$ negativ semidefinit und $\left(\begin{bmatrix} A & \bar{D}\end{bmatrix} \begin{bmatrix} A & \bar{D}\end{bmatrix}^T - \bar{\sigma}_m^2 I\right)^+$ positiv semidefinit ist, sind deshalb notwendigerweise

$$\bar{\sigma}_1^2 I - \boldsymbol{T}\,\boldsymbol{T}^T \quad \text{und} \quad \boldsymbol{T}\,\boldsymbol{T}^T - \bar{\sigma}_m^2 I$$

positiv semidefinit.

(iii) Eine mögliche optimale Wahl ergibt sich aus (3.6). Andere optimale Varianten ergeben sich, wenn σ_1 durch Werte zwischen σ_1 und σ_{m-l} ersetzt wird. Für Y steht dies bereits in [EHM95], wobei dort auf [BGMN92] verwiesen wird.

(iv) Dies entspricht nicht genau dem Vorschlag von Shen [She97], der normierte Links- bzw. Rechtssingulärvektoren von $\partial_{(x,\theta)} F$ in einem Punkt für alle Punkte aus einer gewissen Umgebung verwendete. Bei $l = 1$ und $n - m + q = 2$ ist im aktuellen Punkt $\bar{D}$ zwar ein Linkssingulärvektor zum kleinsten Singulärwert, seine Norm liegt aber im Intervall $\left[\sqrt{\sigma_{m-1}^2 - \sigma_m^2}, \sqrt{\sigma_1^2 - \sigma_m^2}\right]$. Zu diesem Intervall gehört die 1 nicht notwendigerweise. Weiterhin liegt, außer im Fall $\sigma_m = 0$, $\operatorname{im} \bar{V} \begin{bmatrix} 0 & Y_4^T\end{bmatrix}^T = \ker \begin{bmatrix} A & \bar{D}\end{bmatrix}$ nicht im $\operatorname{im} \begin{bmatrix} V^T & 0\end{bmatrix}^T = \operatorname{im} \begin{bmatrix} I_{n+q-l} & 0\end{bmatrix}^T$.

(v) Falls die Berechnung von

$$\left(\sigma_1^2 I_m - A\,A^T\right)^+ \bar{D}$$

durch Lösung eines Gleichungssystems erfolgt, kann auch $X_1 = 0$ überprüft werden. Dieses Gleichungssystem ist nämlich genau dann lösbar, wenn $X_1 = 0$ erfüllt ist. Bei der Berechnung von $\bar{D}^T \left(\sigma_1^2 I_m - A\,A^T\right)^+ \bar{D}$ kann dann eine beliebige Lösung dieses Gleichungssystems verwendet werden, da dieses Produkt davon nicht beeinflußt wird. Bei genügend genauer Kenntnis von σ_1 kann somit

$$I_l - \bar{D}^T \left(\sigma_1^2 I_m - A\,A^T\right)^+ \bar{D}$$

berechnet werden. Analoge Aussagen gelten für

$$\begin{array}{lll} I_l + \bar{D}^T \left(A\,A^T - \sigma_{m-l}^2 I_m\right)^+ \bar{D} & \text{und} & X_3 = 0\,, \\ I_{n-m+q} - \boldsymbol{T} \left(\bar{\sigma}_1^2 I_{n+q} - \begin{bmatrix} A & \bar{D} \end{bmatrix}^T \begin{bmatrix} A & \bar{D} \end{bmatrix}\right)^+ \boldsymbol{T}^T & \text{und} & Y_1 = 0 \qquad \text{und für} \\ I_{n-m+q} + \boldsymbol{T} \left(\begin{bmatrix} A & \bar{D} \end{bmatrix}^T \begin{bmatrix} A & \bar{D} \end{bmatrix} - \bar{\sigma}_m^2 I_{n+q}\right)^+ \boldsymbol{T}^T & \text{und} & Y_3 = 0\,. \end{array}$$

(vi) Bei $l = m$ ist im singulären Punkt $A = 0$. Der optimale Wert von $\bar{\sigma}_1/\bar{\sigma}_m$ ist 1. Dies wird im singulären Punkt genau dann erreicht, wenn $\sigma_1\left(\bar{D}\right) \neq 0$ und $\bar{D}/\sigma_1\left(\bar{D}\right) \in \mathbb{R}^{m\times m}$ eine orthogonale Matrix ist. Aus Satz 3.10 folgt dann die Optimalität der Konditionszahl von $\begin{bmatrix} A & \bar{D} \\ \boldsymbol{T} & \end{bmatrix}$ genau dann, wenn $\boldsymbol{T}\,N_{n+m,m} = 0$ und $\boldsymbol{T}\,M_{n+m,n}/\sigma_1\left(\bar{D}\right) \in \mathbb{R}^{n\times n}$ orthogonal ist. Zum Beispiel kann in jeder Zeile und jeder Spalte der geränderten Matrix jeweils genau eine Eins stehen. Der Fall $l = m$ und $A \neq 0$ wird in der folgenden Proposition 3.12 behandelt.

Proposition 3.12.
Bei $l = m$ ist genau dann $\hat{\sigma}_1 = \hat{\sigma}_{n+m} > 0$, wenn X sich als Produkt von $\operatorname{diag}\left(\sqrt{\hat{\sigma}_1^2 - \sigma_i^2}, i = 1, \ldots, m\right)$ *mit einer orthogonalen Matrix darstellen läßt und $Y = \hat{\sigma}_1\,N_{n+m,n}$, multipliziert mit einer orthogonalen Matrix, ist.*

Beweis. Wegen $\hat{\sigma}_1 \geq \bar{\sigma}_1 \geq \bar{\sigma}_m \geq \hat{\sigma}_{n+m}$ ist $\hat{\sigma}_1 = \hat{\sigma}_{n+m}$ genau dann erfüllt, wenn (3.8) erfüllt ist. Dies entspricht (3.9), vergleiche Bemerkung 3.5, und somit

$$\bar{\sigma}_1^2 I_{n+m} = \begin{bmatrix} \bar{\Sigma}\,\bar{\Sigma}^T & \bar{\Sigma}\,Y \\ Y^T\bar{\Sigma}^T & Y^T Y \end{bmatrix} \quad \text{und} \quad \bar{\sigma}_1^2 I_m = \Sigma\,\Sigma^T + X\,X^T.$$

Dies bedeutet, daß Y nur in den letzten n Zeilen nichtverschwindende Komponenten besitzt, $Y^T Y = \bar{\sigma}_1^2 I_n$ und $X\,X^T = \operatorname{diag}\left(\bar{\sigma}_1^2 - \sigma_i^2, i = 1, \ldots, m\right)$ ist. Für $\bar{\sigma}_1 > \sigma_1$ ergibt sich wegen $\hat{\sigma}_1 = \bar{\sigma}_1$ daraus die Aussage der Proposition. Bei $\bar{\sigma}_1 = \sigma_1$ bedeutet dies, daß die i-ten Zeile von X nur Nullelemente enthält, falls $\sigma_1 = \sigma_i$ ist. Unter Beachtung dieser Beziehung folgt auch in diesem Fall die Aussage der Proposition. □

3.2 Verallgemeinerte inverse Iteration

Damit die Konditionszahl der geränderten Matrix $\begin{bmatrix} A & \bar{D} \\ \boldsymbol{T} & \end{bmatrix}$ nicht viel größer als die minimal erreichbare Konditionszahl σ_1/σ_{m-l} ist, werden hinreichend gute Näherungen von bestimmten Singulärvektoren von A bzw. $\begin{bmatrix} A & \bar{D} \end{bmatrix}$ benötigt. Speziell kann z. B. eine Näherung der in (3.6) bzw. in der Bemerkung 3.11 (iii) vorgeschlagenen Ränderung erfolgen. Dazu ist aber eine hinreichend genaue Kenntnis der zu den jeweils kleinsten Singulärwerten gehörenden Singulärvektoren erforderlich.

Aufbauend auf Stewart [Ste81] schlugen Chan [Cha84, Algorithmus auf S. 746] und Lui [Lui97, Algorithmus auf S. 566] die inverse Vektoriteration zur Bestimmung des kleinsten Singulärwerts und der zugehörigen Singulärvektoren vor. Praktisch ist dies für eine gegebene Matrix C die Anwendung der inversen Iteration zur Bestimmung des kleinsten Eigenwerts und der zugehörigen Eigenvektoren auf die Matrizen C^TC bzw. CC^T. Die Matrizen C^TC bzw. CC^T sind offensichtlich symmetrisch und positiv semidefinit. Somit können die gesuchten Singulärvektoren mittels einer entsprechenden Übertragung der inversen Teilraumiteration bestimmt werden. Die inverse Teilraumiteration zur Berechnung der kleinsten Eigenwerte und der zugehörigen Eigenvektoren symmetrischer regulärer Matrizen schlugen beispielsweise Schwarz/Rutishauser/Stiefel [SRS69, S. 187] vor. Dort wird dieses Verfahren simultane inverse Vektoriteration genannt. Dabei wird die Teilraumiteration, in [SRS69, S. 182ff.] simultane Vektoriteration genannt, auf C^{-1} statt auf die reguläre Matrix C angewendet. Die Matrix C^{-1} wird natürlich nicht explizit berechnet. Stattdessen werden Gleichungssysteme mit der Matrix C gelöst. Die Teilraumiteration selbst wurde von Bauer [Bau57] vorgeschlagen. Er nannte sie stabilisierte Simultaniteration. Die Stabilisierung sichert dabei die Beschränktheit der Matrizenfolge. Für die Stabilisierung schlug Bauer verschiedene Verfahren vor. Bei einer Variante schlug er die parallele Teilraumiteration mit C und mit C^T vor. Dies entspricht der parallelen Berechnung von Links- und Rechtseigenvektoren. Die so erhaltenen Matrizen der genäherten Links- bzw. Rechtseigenvektoren sollen dabei aneinander orthogonalisiert werden. Die so erhaltene Teilraumiteration wird von Bauer [Bau57, § 4] gegenseitig orthogonalisierte Bi-Iteration kurz Bi-Iteration genannt. Zur Orthogonalisierung werden dabei obere Dreiecksmatrizen verwendet. Bauer weist darauf hin, daß dann die Spalten der iterierten Teilraummatrizen gegen Eigenvektoren konvergieren. Für symmetrische positiv definite Matrizen mit einfachen Eigenwerten wird dies von Schwarz/Rutishauser/Stiefel [SRS69, Satz 4.23, S. 182ff.] nachgewiesen. In diesem Fall stimmen die Links- und die Rechtseigenvektoren überein. Deshalb erfolgt nur die Teilraumiteration mit C, wobei in jedem Schritt die Iterationsmatrix orthonormalisiert wird. Dies erfolgt mit der Gram-Schmidt-Orthogonalisierung, von Schwarz/Rutishauser/Stiefel [SRS69] wie auch von Rutishauser [Rut69] Schmidtsches Orthogonalisierungsverfahren genannt. Falls die Stabilisierung auf anderem Wege erfolgt, kann unter vergleichbaren Bedingungen nur gezeigt werden, daß die durch die Spalten aufgespannten Teilräume gegen den durch die entsprechenden Eigenvektoren aufgespannten Raum konvergieren. Rutishauser [Rut69] wendete die Bi-Iteration von Bauer wie in [SRS69] auf symmetrische positiv definite Matrizen an. Weiterhin schlug Rutishauser eine Kombination aus einigen Schritten Teilraumiteration, einer Gram-Schmidt-Orthogonalisierung und einer Ritz-Iteration vor. Dieses Verfahren untersuchte er näher. In [Rut70] steht der dazugehörige ausführliche Algorithmus. Er ist auch für bestimmte symmetrische nicht-definite Matrizen geeignet. Wilkinson [Wil65] verwendete die Variante aus [SRS69] auch für nichtsymmetrische Matrizen. Dieses Vorgehen untersuchte

Stewart [Ste69] näher.

Aus der Literatur ergeben sich noch weitere Methoden zur Bestimmung der zu den kleinsten Singulärwerten gehörenden Singulärvektoren. Van Huffel/Vandewalle/Haegemans [HVH87] untersuchten die partielle Singulärwertzerlegung. Diese ergibt sich aus der Singulärwertzerlegung. Dabei wird ausgenutzt, daß nur die kleinsten Singulärwerte und die zugehörigen Singulärvektoren gesucht sind. Dies führt zu einer Reduzierung des Rechenaufwands. Andere Verfahren entstehen durch Anwendung der Verfahren zur Berechnung der kleinsten Eigenwerte und der zugehörigen Eigenvektoren auf C^TC bzw. CC^T, siehe z.B. [KS88, Kapitel 13] und [SRS69, Kapitel 4]. Bei Verwendung der symmetrischen Matrizen $\begin{bmatrix} 0 & C \\ C^T & 0 \end{bmatrix}$ bzw. $\begin{bmatrix} 0 & C^T \\ C & 0 \end{bmatrix}$ ist zu beachten, daß die kleinsten Singulärwerte von C in der Mitte des Spektrums dieser Matrizen liegen und neben σ_i auch $-\sigma_i$ zum Spektrum gehört, vergleiche z.B. [GVL89, S. 427]. Die erste Eigenschaft spricht gegen die Verwendung des Lanczos-Algorithmus, vergleiche [Pai74, S. 201] bzw. [GVL89, S. 499]. Letztere Eigenschaft stört dagegen die Konvergenz der inversen Teilraumiteration.

In dieser Arbeit soll eine einfache Variante der inversen Teilraumiteration verwendet werden. Diese kann problemlos in die anderen in der Literatur beschriebenen Varianten überführt werden. Die Veränderung zu den erwähnten Arbeiten besteht darin, daß nicht die schlecht konditionierte Matrix A, sondern besser konditionierte geränderte Matrizen verwendet werden. Insbesondere werden die Ergebnisse von [SS97b] für die inverse Iteration auf die inverse Teilraumiteration übertagen.

Zuerst werden in diesem Abschnitt einige später benötigte Größen definiert und Beziehungen hergeleitet. Darauf aufbauend wird im Unterabschnitt 3.2.2 der Algorithmus 3 (`ASVII`) aus [SS97b] auf den Fall der inversen Teilraumiteration übertragen und die Eigenschaften des so erhaltenen Algorithmus untersucht. Speziell wird im Unterabschnitt 3.2.3 die Konvergenz gewisser berechneter Größen gegen die kleinsten Singulärwerte der Ausgangsmatrix betrachtet. Um weitere Eigenschaften nachweisen zu können, ist die Regularität der Ausgangsmatrix hinreichend. Im Unterabschnitt 3.2.4 wird gezeigt, wie relativ einfach aus einer beliebigen gegebenen Matrix durch Ränderung mit Nullraumvektoren eine solche reguläre Matrix gewonnen werden kann. Wenn mit dieser Matrix die inverse Teilraumiteration durchgeführt wird, ergeben sich außerdem weitere Vorteile. Diese reguläre Matrix ist genau dann nicht mit der Ausgangsmatrix identisch, wenn die Ausgangsmatrix nicht regulär ist. In diesem Fall sind die noch zu berechnenden Ränderungsmatrizen von geringerer Dimension, als wenn die inverse Teilraumiteration auf die Ausgangsmatrix angewendet würde. Ein etwas aufwendigerer Algorithmus als im Unterabschnitt 3.2.2 wird im Unterabschnitt 3.2.5 beschrieben und untersucht. Diese besitzt weitere wünschenswerte Eigenschaften. Insbesondere erfolgt nicht nur eine Konvergenz gewisser Größen gegen die von den gesuchten Singulärvektoren aufgespannten Teilräume, sondern unter gewissen Bedingungen auch gegen die Singulärvektoren bzw. -werte selbst. Im Unterabschnitt 3.2.6 werden die erwähnten weiteren Eigenschaften für reguläre Ausgangsmatrizen bzw. für den Algorithmus aus Unterabschnitt 3.2.5 nachgewiesen. Schließlich wurden einige Beispiele mit dem kombinierten Algorithmus aus den Unterabschnitten 3.2.4 und 3.2.5 berechnet. Die dabei erhaltenen numerische Ergebnisse werden im Unterabschnitt 3.2.7 ausgewertet.

3.2.1 Allgemeine Bemerkungen

Es sei $A = \partial_{(x,\theta)} F = \partial_{(x,\theta)} \boldsymbol{F} \in \mathbb{R}^{m\times(n+q-l)}$ in einem gegebenen Punkt (x, λ, α) in der Umgebung des singulären Punktes $(x^*, \lambda^*, \alpha^*)$. Dabei sei wieder $\sigma_{m-l+1} < \sigma_{m-l}$, vergleiche Bemerkung 3.3 (iv), woraus $\operatorname{rank}(A) \geq m - l$ folgt. Auch hier wird der Fall $l = m$ extra behandelt. Deshalb wird in diesem Abschnitt $l < m$ vorausgesetzt. Aufbauend auf Definition 3.2 werden folgende weitere Bezeichnungen eingeführt.

Definition 3.13. Es sei

$$T_0 := M^T_{n+q,n+q-l} \boldsymbol{T}^T \in \mathbb{R}^{(n+q-l)\times(n-m+q)} \quad \text{und} \quad T_1 := N^T_{n+q,l} \boldsymbol{T}^T \in \mathbb{R}^{l\times(n-m+q)},$$

d. h. $\boldsymbol{T} = \begin{bmatrix} T_0^T & T_1^T \end{bmatrix}$. Weiterhin seien

$$\begin{aligned} A &= \begin{bmatrix} U_0 & U_4 \end{bmatrix} \left[\begin{array}{c|c} \Sigma_0 & 0 \\ \hline 0 & \Sigma_4 \end{array}\right] \begin{bmatrix} V_0 & V_4 \end{bmatrix}^T, \\ \Sigma_0 &= \operatorname{diag}(\Sigma_i, i = 1,2,3) = \operatorname{diag}(\sigma_i, \sigma_1 \geq \sigma_i \geq \sigma_{m-l}), \\ \begin{bmatrix} X_0 \\ X_4 \end{bmatrix} &= U^T \bar{D} = \begin{bmatrix} U_0^T \bar{D} \\ U_4^T \bar{D} \end{bmatrix}, \quad \begin{bmatrix} Z_0 \\ Z_4 \end{bmatrix} = V^T T_0 = \begin{bmatrix} V_0^T T_0 \\ V_4^T T_0 \end{bmatrix}, \qquad \text{so daß} \\ \bar{D} &= U_0 X_0 + U_4 X_4 \quad \text{und} \quad T_0 = V_0 Z_0 + V_4 Z_4 \end{aligned} \tag{3.17}$$

gilt. Schließlich seien $X_s = U_s^T \bar{D}$ und $Z_s = V_s^T T_0$ die Matrizen der Zeilen von $U^T \bar{D}$ bzw. $V^T T_0$, die zum Singulärwert Null gehören und

$$B := B(\bar{D}, T_0, T_1) := \left[\begin{array}{c|c} A & \bar{D} \\ \hline T_0^T & T_1^T \end{array}\right] = \begin{bmatrix} A & \bar{D} \\ \boldsymbol{T} & \end{bmatrix} \in \mathbb{R}^{(n+q)\times(n+q)}. \tag{3.18}$$

Bemerkung 3.14.

(i) Wegen $\sigma_{m-l+1} < \sigma_{m-l}$ ist $\Sigma_0 \in \mathbb{R}^{(m-l)\times(m-l)}$ und $\Sigma_4 \in \mathbb{R}^{l\times(n-m+q)}$.

(ii) Bei $\sigma_m \neq 0$ ist X_s nicht vorhanden, während bei $\sigma_{n+q-l} \neq 0$ die Matrix Z_s nicht existiert. Dagegen gelten bei $\sigma_{m-l+1} = 0$ mit $l > 0$ die Gleichungen $X_4 = X_s$ und $Z_4 = Z_s$.

(iii) Bei $l = 0$ ist $T_0 = \boldsymbol{T}^T$, während T_1, U_4, Σ_4, X, X_0, X_4 und X_s nicht existieren und $B^T = \begin{bmatrix} A^T & \boldsymbol{T}^T \end{bmatrix}$. Dieser Spezialfall wird jeweils gesondert betrachtet.

(iv) Bei $l = m$ ist $A = U_4 \Sigma_4 V_4$, $\bar{D} = U_4 X_4$ und $T_0 = V_4 Z_4$.

Mit diesen Bezeichnungen gilt:

Lemma 3.15.
Die folgenden Aussagen sind äquivalent:

(i) $B(\bar{D}, T_0, T_1)$ *ist regulär.*

(ii) $\operatorname{rank}\begin{bmatrix} A & \bar{D} \end{bmatrix} = m, \quad \operatorname{rank}\begin{bmatrix} T_0^T & T_1^T \end{bmatrix} = \operatorname{rank}(\boldsymbol{T}) = n - m + q \quad$ *und*

$$\operatorname{im}\begin{bmatrix} T_0 \\ T_1 \end{bmatrix} \cap \operatorname{im}\begin{bmatrix} A^T \\ \bar{D}^T \end{bmatrix} = \operatorname{im} \boldsymbol{T}^T \cap (\ker \begin{bmatrix} A & \bar{D} \end{bmatrix})^{\perp} = \{0_{\mathbb{R}^{n+q}}\}$$

(iii) $\operatorname{rank}\begin{bmatrix} A \\ T_0^T \end{bmatrix} = n + q - l, \quad \operatorname{rank}\begin{bmatrix} \bar{D} \\ T_1^T \end{bmatrix} = l \quad$ *und*

$$\operatorname{im}\begin{bmatrix} \bar{D} \\ T_1^T \end{bmatrix} \cap \operatorname{im}\begin{bmatrix} A \\ T_0^T \end{bmatrix} = \operatorname{im}\begin{bmatrix} \bar{D} \\ T_1^T \end{bmatrix} \cap (\ker \begin{bmatrix} A^T & T_0 \end{bmatrix})^{\perp} = \{0_{\mathbb{R}^{n+q}}\}$$

(iv) Im Fall $\operatorname{rank}(A) = m - l$ *sind* $X_4 = U_4^T \bar{D}$ *und* $Z_4 = V_4^T T_0$ *regulär.*

Daraus folgt die Zeilenregularität von X_s *und* Z_s.

Bemerkung 3.16.

(i) Im Fall $\operatorname{rank}(A) = m - l$ ist $X_s = X_4$ und $Z_s = Z_4$. In diesem Fall ist wegen Lemma 3.15 (iv) die Zeilenregularität von X_s und Z_s nicht nur notwendig, sondern auch hinreichend für die Regularität von $B(\bar{D}, T_0, T_1)$.

(ii) Die Äquivalenz von (i) und (iv) wurde bereits von Keller [Kel77, S. 363, Lemma 2.8] für lineare Operatoren in Banachräumen bei $m = n + q - l$ in einer gleichwertigen Form erwähnt.

(iii) Im Fall $l = 0$ wird in Lemma 3.15 (ii) $\begin{bmatrix} A & \bar{D} \end{bmatrix}$ zu A, $\begin{bmatrix} T_0^T & T_1^T \end{bmatrix}$ zu $T_0^T = \boldsymbol{T}$, Lemma 3.15 (iii) wird zu $\operatorname{rank}\begin{bmatrix} A^T & T_0 \end{bmatrix} = n + q$, was unmittelbar Lemma 3.15 (i) entspricht. In Lemma 3.15 (iv) wird dann nur die Regularität von Z_4 verlangt.

Beweis. Die Äquivalenz von (i) und (ii) folgt aus einfachen Rangbetrachtungen. Falls statt $B(\bar{D}, T_0, T_1)$ die Matrix $B(\bar{D}, T_0, T_1)^T$ betrachtet wird, ergibt diese Äquivalenz diejenige von (i) und (iii). Im Fall $\operatorname{rank}(A) = m - l$ ist genau dann $\operatorname{rank}\begin{bmatrix} A & \bar{D} \end{bmatrix} = m$, wenn $X_4 = U_4^T \bar{D}$ regulär ist. Dann gilt $\ker \begin{bmatrix} A & \bar{D} \end{bmatrix} = \operatorname{im} \begin{bmatrix} V_4^T & 0 \end{bmatrix}^T$. Daraus folgt die Äquivalenz von (i) und (iv). Wegen $\operatorname{rank}\begin{bmatrix} A & \bar{D} \end{bmatrix} = m$ folgt die Zeilenregularität von X_s und wegen $\operatorname{rank}\begin{bmatrix} A^T & T_0 \end{bmatrix} = n + q - l$ die Zeilenregularität von Z_s. □

Bemerkung 3.17. Lemma 3.15 gilt offensichtlich auch im Fall $l = m$, wobei $n - m + q = n$, $n + q = n + m$, $n + q - l = n$, $m - l = 0$, $X_4 = X$ und $Z_4 = Z$ ist.

Der Winkel zwischen den Teilräumen $\operatorname{im} U_4$, $\operatorname{im} \bar{D} \subseteq \mathbb{R}^m$ ist im Fall $l > 0$ definiert als

(3.19)

$$\begin{aligned} \xi &:= \measuredangle(\operatorname{im} \bar{D}, \operatorname{im} U_4) := \max\{\min\{\measuredangle(d, u) : u \in \operatorname{im} U_4\} : d \in \operatorname{im} \bar{D}\} \\ &:= \max\left\{ \min\left\{ \begin{array}{ll} \pi/2 & \text{if} \quad d^T u = 0, \\ \arccos\left[\dfrac{|d^T u|}{\|d\| \, \|u\|}\right] \in [0, \pi/2) & \text{if} \quad d^T u \neq 0. \end{array} : u \in \operatorname{im} U_4 \right\} : d \in \operatorname{im} \bar{D} \right\}. \end{aligned}$$

Dabei sei die Norm hier und im Folgenden die Euklidische bzw. bei Matrizen die Spektralnorm. Analog sei

$$\zeta := \measuredangle(\operatorname{im} T_0, \operatorname{im} V_4) .$$

Für orthonormale Matrizen $\bar{D}$ und T_0, d. h. $\bar{D}^T\bar{D} = I_l$ und $T_0^T T_0 = I_{n-m+q}$, und reguläre Matrizen X_4 und Z_4 gilt dann

$$\tan\xi = \|X_0\|\,\|(X_4)^{-1}\| = \|X_0(X_4)^{-1}\|, \quad \tan\zeta = \|Z_0\|\,\|(Z_4)^{-1}\| = \|Z_0(Z_4)^{-1}\|, \tag{3.20}$$

vergleiche Golub/Van Loan [GVL89, § 2.6.3 und § 12.4.3] und Rheinboldt [Rhe93].

Lemma 3.18.
Falls X_4 und Z_4 regulär aber die Spalten von $\bar{D}$ bzw. T_0 nicht notwendigerweise orthonormal sind, ist ebenfalls

$$\tan\xi = \|X_0(X_4)^{-1}\|, \quad \tan\zeta = \|Z_0(Z_4)^{-1}\|.$$

Bemerkung 3.19. Im Fall $l = 0$ existiert der Winkel ξ nicht, weshalb in diesem Fall nur die Aussagen über ζ zutreffen.

Beweis. Aus der Regularität von X_4 folgt die Spaltenregularität von $\bar{D} = U\,X$ und damit die positive Definitheit von $\bar{D}^T\bar{D}$. Offensichtlich gilt $\left(\bar{D}^T\bar{D}\right)^{-\frac{1}{2}}\bar{D}^T\bar{D}\left(\bar{D}^T\bar{D}\right)^{-\frac{1}{2}} = I_l$, woraus aus (3.20)

$$\tan\xi = \left\| X_0\left(\bar{D}^T\bar{D}\right)^{-\frac{1}{2}}\left(X_4\left(\bar{D}^T\bar{D}\right)^{-\frac{1}{2}}\right)^{-1}\right\| = \|X_0(X_4)^{-1}\|$$

folgt. Analog läßt sich $\tan\zeta = \|Z_0(Z_4)^{-1}\|$ zeigen. □

Die Teilräume $\operatorname{im}\bar{D}$ bzw. $\operatorname{im} T_0$ sollen nun durch Lösung der Systeme

$$B(\bar{D}, T_0, T_1)\,\tilde{\boldsymbol{T}}^T = N_{n+q,n-m+q} \quad \text{und} \quad B^T(\bar{D}, T_0, T_1)\,\hat{D} = N_{n+q,l}\,, \tag{3.21}$$

d. h.

$$\begin{bmatrix} A & \bar{D} \\ T_0^T & T_1^T \end{bmatrix}\begin{bmatrix} \tilde{T}_0 \\ \tilde{J} \end{bmatrix} = \begin{bmatrix} 0 \\ I_{n-m+q} \end{bmatrix} \quad \text{und} \quad \begin{bmatrix} A^T & T_0 \\ \bar{D}^T & T_1 \end{bmatrix}\begin{bmatrix} \tilde{D} \\ \tilde{K} \end{bmatrix} = \begin{bmatrix} 0 \\ I_l \end{bmatrix}, \tag{3.22}$$

verbessert, d. h., $\operatorname{im} U_4$ bzw. $\operatorname{im} V_4$ angenähert werden. Diese Verbesserung wird durch das folgende Lemma beschrieben.

Lemma 3.20.
Es seien $0 < l < m$, $\operatorname{rank}(A) \geq m - l$ und $\bar{D}$ und $\boldsymbol{T}$ so gewählt, daß $B(\bar{D}, T_0, T_1)$ regulär ist. Dann sind die Systeme (3.22) eindeutig lösbar, $\tilde{\boldsymbol{T}}^T$ und $\hat{D}$ sind spaltenregulär, $\tilde{J} = \tilde{K}^T$ und

$$\tilde{J} = \tilde{K}^T = 0 \quad \Longleftrightarrow \quad \operatorname{rank}(A) = m - l\,.$$

Die neuen Matrizen $\tilde{T}_0$ und $\tilde{D}$ besitzen folgende Eigenschaften:

(i) Im Fall $\operatorname{rank}(A) = m-l$, *d. h.* $\sigma_{m-l+1} = 0$, *d. h.* $\kappa := \sigma_{m-l+1}/\sigma_{m-l} = 0$, *ist* $X_4 = U_4^T \bar{D}$ *und* $Z_4 = V_4^T T_0$ *regulär und*

$$\tilde{D} = U_4 \, (X_4)^{-T}, \quad \tilde{T}_0 = V_4 \, (Z_4)^{-T}. \tag{3.23}$$

(ii) Mit $0 \le \kappa = \sigma_{m-l+1}/\sigma_{m-l} \le 1$ *gilt bei* $\tilde{\zeta} := \measuredangle(\operatorname{im} \tilde{T}_0, \operatorname{im} V_4)$ *und* $\tilde{\xi} := \measuredangle(\operatorname{im} \tilde{D}, \operatorname{im} U_4)$*:*

(a) Wenn $X_4 = U_4^T \bar{D}$ *regulär ist, ist* $\tilde{Z}_4 = V_4^T \tilde{T}_0$ *ebenfalls regulär,*

$$\tan \tilde{\zeta} \le \kappa \tan \xi \tag{3.24}$$

und $\operatorname{im} \begin{bmatrix} V_s^T & 0 \end{bmatrix}^T \subseteq \operatorname{im} \tilde{\boldsymbol{T}}^T$.

(b) Wenn $Z_4 = V_4^T T_0$ *regulär ist, ist* $\tilde{X}_4 = U_4^T \tilde{D}$ *ebenfalls regulär,*

$$\tan \tilde{\xi} \le \kappa \tan \zeta \tag{3.25}$$

und $\operatorname{im} \begin{bmatrix} U_s^T & 0 \end{bmatrix}^T \subseteq \operatorname{im} \hat{D}$.

Bemerkung 3.21.

(i) Im Fall $l = 0$ hat (3.21) weiterhin die Gestalt $B \, \tilde{\boldsymbol{T}}^T = N_{n+q,n-m+q}$, was jedoch bei (3.22) $\begin{bmatrix} A \\ \boldsymbol{T} \end{bmatrix} \tilde{T}_0 = \begin{bmatrix} 0 \\ I_{n-m+q} \end{bmatrix}$ entspricht.

(ii) Analog zu Lemma 3.20 ergibt sich dann die Spaltenregularität von $\tilde{\boldsymbol{T}}^T = \tilde{T}_0$, die Regularität von Z_4 und $\tilde{T}_0 = V_4 \, (Z_4)^{-T}$.

Beweis. Offensichtlich gilt, daß $\tilde{\boldsymbol{T}}^T$ und $\hat{D}$ spaltenregulär sind und

$$\begin{aligned} \tilde{J} &= N_{n+q,l}^T \tilde{\boldsymbol{T}}^T = N_{n+q,l}^T B^{-1} N_{n+q,n-m+q} = (N_{n+q,n-m+q}^T B^{-T} N_{n+q,l})^T = (N_{n+q,n-m+q}^T \hat{D})^T \\ &= \tilde{K}^T. \end{aligned}$$

Teil (i): Wegen Lemma 3.15 (iv) sind im Fall $\operatorname{rank}(A) = m - l$ die Matrizen X_4 und Z_4 regulär. Die Multiplikation des oberen Teils

$$A \, \tilde{T}_0 + \bar{D} \, \tilde{J} = 0 \tag{3.26}$$

des linken Systems von (3.22) mit U_4^T liefert $U_4^T A \, \tilde{T}_0 + U_4^T \bar{D} \, \tilde{J} = 0 + X_4 \, \tilde{J} = 0$ und somit $\tilde{J} = 0$, woraus $A\tilde{T}_0 = 0$ und damit $\tilde{T}_0 = V_4 \, C$ folgt. Wenn dies in den unteren Teil des linken Systems von (3.22) eingesetzt wird, ergibt sich $T_0^T \tilde{T}_0 + T_1^T \tilde{J} = T_0^T V_4 \, C + 0 = I_{n-m+q}$ und damit $\tilde{T}_0 = V_4 \, (Z_4)^{-T}$. Der andere Teil von (3.23) kann analog aus der rechten Gleichung von (3.22) gefolgert werden.
Teil (ii): Unter Nutzung der Singulärwertzerlegung von A und der Zerlegung von $\bar{D}$ gemäß (3.17) ergibt sich aus (3.26)

$$\begin{aligned} 0 = A \, \tilde{T}_0 + \bar{D} \, \tilde{J} &= U_0 \, \Sigma_0 \, V_0^T \tilde{T}_0 + U_4 \, \Sigma_4 \, V_4^T \tilde{T}_0 + U_0 \, U_0^T \bar{D} \, \tilde{J} + U_4 \, U_4^T \bar{D} \, \tilde{J} \\ &= U_0 \, (\Sigma_0 \, \tilde{Z}_0 + X_0 \, \tilde{J}) + U_4 \, (\Sigma_4 \, \tilde{Z}_4 + X_4 \, \tilde{J}) \end{aligned}$$

mit $\tilde{T}_0 = V_0 \tilde{Z}_0 + V_4 \tilde{Z}_4$. Daraus folgt

$$\tilde{Z}_0 = -\Sigma_0^{-1} X_0 \tilde{J} \quad \text{und} \quad \tilde{J} = -(X_4)^{-1} \Sigma_4 \tilde{Z}_4. \tag{3.27}$$

Aus der Zeilenregularität von $\tilde{\boldsymbol{T}}$ folgt damit die Spaltenregularität und damit die Regularität von $\tilde{Z}_4$ und

$$\tan \tilde{\zeta} = \|\tilde{Z}_0 (\tilde{Z}_4)^{-1}\| = \|\Sigma_0^{-1} X_0 (X_4)^{-1} \Sigma_4\| \leq \|\Sigma_0^{-1}\| \, \|X_0 (X_4)^{-1}\| \, \|\Sigma_4\| = \kappa \tan \xi \,. \tag{3.28}$$

Bei $\operatorname{rank}(A) > m - l$, d. h. $\sigma_{m-l+1} > 0$, ergibt sich daraus $\tilde{J} \neq 0$. Aus (3.27) kann dann gefolgert werden, daß die letzten $n+q-l-\operatorname{rank}(A)$ Spalten von $\tilde{\boldsymbol{T}}^T \begin{pmatrix} \tilde{Z}_4 \end{pmatrix}^{-1}$ gerade $\begin{bmatrix} V_s^T & 0 \end{bmatrix}^T$ sind. Daraus ergibt sich im $\begin{bmatrix} V_s^T & 0 \end{bmatrix}^T \subseteq \operatorname{im} \tilde{\boldsymbol{T}}^T$. Teil (b) kann analog mittels des rechten Systems aus (3.22) nachgewiesen werden. □

Bemerkung 3.22.

(i) Die Winkel $\tilde{\xi}$ zwischen im $\tilde{D}$ und im U_4 bzw. $\tilde{\zeta}$ zwischen im $\tilde{T}_0$ und im V_4 hängen nur von den in den ersten Zeilen von B^T bzw. B stehenden Ränderungsmatrizen T_0 bzw. $\bar{D}$ ab. Die unteren Zeilen von B^T bzw. B beeinflussen diese Winkel nicht. Sie sichern lediglich die Regularität von B^T bzw. B und damit die eindeutige Lösbarkeit von (3.22). Überhaupt ergibt sich aus dem Beweis, daß die Aussagen im Lemma 3.20 (ii) unabhängig vom unteren Block der Gleichungssysteme in (3.22) sind. Neben der Regularität von B bzw. B^T sollen diese unteren Blöcke lediglich die Spaltenregularität von $\tilde{\boldsymbol{T}}^T$ und $\hat{D}$ und damit $\ker \begin{bmatrix} A & \bar{D} \end{bmatrix} = \operatorname{im} \tilde{\boldsymbol{T}}^T$ und $\ker \begin{bmatrix} A^T & T_0 \end{bmatrix} = \operatorname{im} \hat{D}$ sichern.

(ii) Im Fall $l = m$ ist offensichtlich $\xi = \zeta = 0$ und somit $\tan \xi = \tan \zeta = 0$, während die Darstellung (3.20) nicht existiert. Lemma 3.20 gilt analog, wobei (3.23) zu $\tilde{D} = (\bar{D})^{-T}$ und $\tilde{T}_0 = (T_0)^{-T}$ wird und (3.24) und (3.25) für beliebige Werte von κ trivial sind, obwohl κ in diesem Fall nicht definiert ist.

Obiges Lemma 3.20 zeigt, daß im Fall $\operatorname{rank}(A) = m - l$ bereits ein Schritt zur Bestimmung von im U_4 und im V_4 ausreicht. Dies ist aus der Verzweigungstheorie bekannt. Andernfalls erfolgt die Annäherung an diese Teilräume, beschrieben durch den Tangens, durch die Iteration (3.22) mit dem Faktor κ. Bei $\sigma_{m-l} > \sigma_{m-l+1}$ ist er kleiner als 1. Dies scheint neu zu sein und ist eine Verallgemeinerung der entsprechenden Aussage für Vektoren aus [SS97b, S. 6]. Auch in diesem Fall reicht bereits ein Schritt zur Bestimmung von im U_s und im V_s aus.

Für reguläre Matrizen A und $l > 0$ kann Lemma 3.20 (ii) verallgemeinert werden. Dies wird für bestimmte Konvergenzuntersuchungen benötigt.

Lemma 3.23.
Es seien $A \in \mathbb{R}^{m \times m}$ und B regulär, $0 < l < m$, $X_j \in \mathbb{R}^{l \times l}$ sei eine reguläre Teilmatrix von $X = U^T \bar{D}$, $X_i \in \mathbb{R}^{(m-l) \times l}$ bestehe aus den restlichen Zeilen von X, $\Sigma_j \in \mathbb{R}^{l \times l}$, $\Sigma_i \in \mathbb{R}^{(m-l) \times (m-l)}$, $\tilde{Z}_j \in \mathbb{R}^{l \times l}$ und $\tilde{Z}_i \in \mathbb{R}^{(m-l) \times l}$ seien die entsprechenden Teilmatrizen von Σ und von $V^T \tilde{T}_0$. Dann ist auch $\tilde{Z}_j$ regulär.

Beweis. Die Übertragung von (3.27) auf den hier betrachteten Fall ergibt

$$\tilde{Z}_i = -\Sigma_i^{-1} X_i \tilde{J} \quad \text{und} \quad \tilde{J} = -(X_j)^{-1} \Sigma_j \tilde{Z}_j. \tag{3.29}$$

Aus der Zeilenregularität von $\tilde{\boldsymbol{T}}$ folgt damit die Regularität von $\tilde{Z}_j$. □

Lemma 3.24.
Es seien A, B und X_4 regulär, $0 < l < m$, $\sigma_1 > \sigma_{m-l}$ und $X_3 \neq 0$. Die reguläre Matrix X_j bestehe aus den linear unabhängigen Zeilen von X_3 und Zeilen von X_4. Dann gilt für die Σ_1 entsprechenden Matrizen $X_1 = U_1^T \bar{D}$ und $\tilde{Z}_1 = V_1^T \tilde{T}_0$ und die X_j entsprechende Matrix $\tilde{Z}_j$ die Beziehung

$$\left\| \tilde{Z}_1 \left(\tilde{Z}_j\right)^{-1} \right\| \leq \frac{\sigma_{m-l}}{\sigma_1} \left\| X_1 \left(X_j\right)^{-1} \right\| . \tag{3.30}$$

Falls sogar Σ_2 existiert, vergleiche Bemerkung 3.3 (ii), gilt für die Σ_2 entsprechenden Matrizen $X_2 = U_2^T \bar{D}$ und $\tilde{Z}_2 = V_2^T \tilde{T}_0$

$$\left\| \tilde{Z}_2 \left(\tilde{Z}_j\right)^{-1} \right\| \leq \frac{\sigma_{m-l}}{\min\{\sigma_j : \sigma_j > \sigma_{m-l}\}} \left\| X_2 \left(X_j\right)^{-1} \right\| . \tag{3.31}$$

Bemerkung 3.25. Da $X_3 \neq 0$ ist, existieren linear unabhängige Zeilen von X_3. Die übrigen Zeilen von X_3 sind Linearkombinationen dieser Zeilen. Da X_4 regulär ist, existiert X_j.

Beweis. Wegen Lemma 3.23 ist $\tilde{Z}_j$ regulär. X_1, Σ_1, $\tilde{Z}_1$ und, falls vorhanden, X_2, Σ_2 bzw. $\tilde{Z}_2$ sind dann Teilmatrizen von X_i, Σ_i bzw. $\tilde{Z}_i$ aus Lemma 3.23. Aus (3.29) ergibt sich dann

$$\tilde{Z}_1 = \Sigma_1^{-1} X_1 \left(X_j\right)^{-1} \Sigma_j \tilde{Z}_j \quad \text{und, falls vorhanden,} \quad \tilde{Z}_2 = \Sigma_2^{-1} X_2 \left(X_j\right)^{-1} \Sigma_j \tilde{Z}_j$$

und damit

$$\tilde{Z}_1 \left(\tilde{Z}_j\right)^{-1} = \Sigma_1^{-1} X_1 \left(X_j\right)^{-1} \Sigma_j \quad \text{und, falls vorhanden,} \quad \tilde{Z}_2 \left(\tilde{Z}_j\right)^{-1} = \Sigma_2^{-1} X_2 \left(X_j\right)^{-1} \Sigma_j .$$

Daraus folgt analog zu (3.28) unter Beachtung von $\|\Sigma_j\| = \sigma_{m-l}$, $\left\|\Sigma_1^{-1}\right\| = (\sigma_1)^{-1}$ und, falls vorhanden, $\left\|\Sigma_2^{-1}\right\| = \left(\min\{\sigma_j : \sigma_j > \sigma_{m-l}\}\right)^{-1}$ (3.30) und (3.31). □

Bemerkung 3.26. Lemma 3.23 und Lemma 3.24 gelten offensichtlich auch, wenn X und Z miteinander vertauscht und die anderen Bezeichnungen entsprechend transformiert werden. Insbesondere wird Σ durch Σ^T ersetzt.

3.2.2 Ein einfacher Algorithmus

In Anlehnung an Algorithmus 3 (`ASVII`) aus [SS97b] soll ein Algorithmus zur alternierenden inversen Teilraumiteration für Singulärwerte vorgestellt und untersucht werden. Dabei wird insbesondere die Konvergenz der Teilräume betrachtet. Im folgenden Unterabschnitt 3.2.3 wird dann gezeigt, daß die Singulärwerte von speziellen Matrizen gegen die Singulärwerte von Σ_4 konvergieren.

Die Verallgemeinerung vom Algorithmus 3 aus [SS97b] ist unter Beachtung von (3.21) und (3.22) der folgende Algorithmus. Dabei wurden u. a. die Indizes verändert. Um diesen Algorithmus sinnvoll ausführen zu können, wird $0 < l < m$ vorausgesetzt.

Algorithmus 3.27.

Initialisierung:

S0: Wähle $\bar{D}^{(0)} \in \mathbb{R}^{m\times l}$ so, daß die Spalten von $\bar{D}^{(0)}$ orthonormal sind und $T_0^{(0)} \in \mathbb{R}^{(n+q-l)\times(n-m+q)}$, $K^{(0)} \in \mathbb{R}^{(n-m+q)\times l}$, setze $i := 0$

Iteration: Schritt $i \to i+1$

solange { das Abbruchkriterium nicht erfüllt ist, } **führe folgende Schritte aus**

S1.1: Löse $\begin{bmatrix} A & \bar{D}^{(i)} \\ (T_0^{(i)})^T & K^{(i)} \end{bmatrix} \begin{bmatrix} \tilde{T}_0^{(i+1)} \\ \tilde{J}^{(i+1)} \end{bmatrix} = \begin{bmatrix} 0 \\ I_{n-m+q} \end{bmatrix}$ für $\left(\tilde{\boldsymbol{T}}^{(i+1)}\right)^T := \begin{bmatrix} \tilde{T}_0^{(i+1)} \\ \tilde{J}^{(i+1)} \end{bmatrix}$, $\tilde{\boldsymbol{T}}^{(i+1)} \in \mathbb{R}^{(n-m+q)\times(n+q)}$

S1.2: Berechne $T_0^{(i+1)} \in \mathbb{R}^{(n+q-l)\times(n-m+q)}$, $R_1^{(i+1)} \in \mathbb{R}^{(n-m+q)\times(n-m+q)}$ aus der QR-Zerlegung $\tilde{T}_0^{(i+1)} = T_0^{(i+1)} R_1^{(i+1)}$, setze $J^{(i+1)} := \tilde{J}^{(i+1)} \left(R_1^{(i+1)}\right)^{-1}$

S2.1: Löse $\begin{bmatrix} A^T & T_0^{(i+1)} \\ (\bar{D}^{(i)})^T & J^{(i+1)} \end{bmatrix} \begin{bmatrix} \tilde{D}^{(i+1)} \\ \tilde{K}^{(i+1)} \end{bmatrix} = \begin{bmatrix} 0 \\ I_l \end{bmatrix}$ für $\hat{D}^{(i+1)} := \begin{bmatrix} \tilde{D}^{(i+1)} \\ \tilde{K}^{(i+1)} \end{bmatrix}$, $\hat{D}^{(i+1)} \in \mathbb{R}^{(n+q)\times l}$

S2.2: Berechne $\bar{D}^{(i+1)} \in \mathbb{R}^{m\times l}$, $R_2^{(i+1)} \in \mathbb{R}^{l\times l}$ aus der QR-Zerlegung $\tilde{D}^{(i+1)} = \bar{D}^{(i+1)} R_2^{(i+1)}$, setze $K^{(i+1)} := \tilde{K}^{(i+1)} \left(R_2^{(i+1)}\right)^{-1}$

S3: Setze $i := i+1$

Ende

Bemerkung 3.28.

(i) Im Fall $l = 0$ wird der Algorithmus zu

Initialisierung:

S0: Wähle $\boldsymbol{T}^{(0)} \in \mathbb{R}^{(n-m+q)\times(n+q)}$

Berechnung:

S1.1: Löse $\begin{bmatrix} A \\ \boldsymbol{T}^{(0)} \end{bmatrix} \left(\tilde{\boldsymbol{T}}^{(1)}\right)^T = \begin{bmatrix} 0 \\ I_{n-m+q} \end{bmatrix}$ für $\tilde{\boldsymbol{T}}^{(1)} \in \mathbb{R}^{(n-m+q)\times(n+q)}$

S1.2: Berechne $\boldsymbol{T}^{(1)} \in \mathbb{R}^{(n-m+q)\times(n+q)}$, $R_1^{(1)} \in \mathbb{R}^{(n-m+q)\times(n-m+q)}$ aus der QR-Zerlegung $\left(\tilde{\boldsymbol{T}}^{(1)}\right)^T = \left(\boldsymbol{T}^{(1)}\right)^T R_1^{(1)}$

Ende

(ii) Damit sind die Zeilen von $\boldsymbol{T}^{(1)}$ wegen Bemerkung 3.21 eine orthonormale Basis von $\ker A$.

(iii) Im Fall $l = m$ ist bereits $\operatorname{im} \bar{D}^{(0)} = \operatorname{im} U_4 = \operatorname{im} U$, weshalb diese Iteration keine Verbesserung bringt.

(iv) Deshalb brauchen die Fälle $l = 0$ und $l = m$ in diesem und den folgenden Unterabschnitt 3.2.3 nicht betrachtet werden.

Der folgende Satz 3.29 beschreibt das Verhalten vom Algorithmus 3.27.

Satz 3.29.
Es seien $0 < l < m$, $\operatorname{rank}(A) \geq m - l$, $\kappa = \sigma_{m-l+1}/\sigma_{m-l} < 1$ *und die Starträndерungen* $\bar{D}^{(0)}$, $T_0^{(0)}$ *und* $K^{(0)}$ *seien so gewählt, daß die Matrizen*

$$X_4^{(0)} = U_4^T \bar{D}^{(0)} \quad \textit{und} \quad B_0 := B(\bar{D}^{(0)}, T_0^{(0)}, (K^{(0)})^T) \quad \textit{regulär sind.} \tag{3.32}$$

Dann ist Algorithmus 3.27 wohldefiniert, d. h., die Matrizen

$$B_{2i} := B(\bar{D}^{(i)}, T_0^{(i)}, (K^{(i)})^T), \quad B_{2i+1} := B(\bar{D}^{(i)}, T_0^{(i+1)}, J^{(i+1)}) \tag{3.33}$$

sind regulär, die Matrizen $\tilde{T}_0^{(i+1)}$ *und* $\tilde{D}^{(i+1)}$ *sind spaltenregulär und*

$$\tan \zeta_{i+1} \leq \kappa \tan \xi_i, \quad \tan \xi_{i+1} \leq \kappa \tan \zeta_{i+1} \quad (i = 0, 1, \ldots), \tag{3.34}$$

wobei ζ_i *und* ξ_i *die Winkel* $\zeta_i := \measuredangle(\operatorname{im} T_0^{(i)}, \operatorname{im} V_4)$ *und* $\xi_i := \measuredangle(\operatorname{im} \bar{D}^{(i)}, \operatorname{im} U_4)$ *bezeichnen.*

Außerdem gilt genau dann $J^{(1)} = (K^{(1)})^T = 0$, *wenn* $\operatorname{rank}(A) = m - l$. *In diesem Fall ist* $J^{(i+1)} = (K^{(i+1)})^T = 0$, $T_0^{(i+1)} = V_4 \left(R_1^{(1)} \left(Z_4^{(0)} \right)^T \right)^{-1}$ *und* $\bar{D}^{(i+1)} = U_4 \left(R_2^{(1)} \left(X_4^{(0)} \right)^T \right)^{-1}$ *für alle* $i \geq 0$ *mit orthogonalen Matrizen* $R_1^{(1)} \left(Z_4^{(0)} \right)^T$ *und* $R_2^{(1)} \left(X_4^{(0)} \right)^T$.

In jedem Fall ist $\operatorname{im} \begin{bmatrix} V_s^T & 0 \end{bmatrix}^T \subseteq \operatorname{im} \left(\tilde{\boldsymbol{T}}^{(i+1)} \right)^T = \operatorname{im} \begin{bmatrix} (T_0^{(i+1)})^T & (J^{(i+1)})^T \end{bmatrix}^T$ *und* $\operatorname{im} \begin{bmatrix} U_s^T & 0 \end{bmatrix}^T \subseteq \operatorname{im} \hat{D}^{(i+1)} = \operatorname{im} \begin{bmatrix} (\bar{D}^{(i+1)})^T & (K^{(i+1)})^T \end{bmatrix}^T$.

Beweis. Aus der Regularität von $X_4^{(0)}$ folgt die Spaltenregularität von $\bar{D}^{(0)}$. Daraus ergibt sich zusammen mit der Regularität von B_0 die Zeilenregularität von $\tilde{\boldsymbol{T}}^{(1)}$ und die Spaltenregularität von $\tilde{T}_0^{(1)}$. Somit ist $\begin{bmatrix} T_0^{(1)} \\ J^{(1)} \end{bmatrix} = \begin{bmatrix} \tilde{T}_0^{(1)} \\ \tilde{J}^{(1)} \end{bmatrix} \left(R_1^{(1)} \right)^{-1}$ definiert und erfüllt $A\, T_0^{(1)} + \bar{D}^{(0)} J^{(1)} = 0$. Da $\begin{bmatrix} A & \bar{D}^{(0)} \end{bmatrix}$ zeilenregulär ist, bedeutet dies $\ker \begin{bmatrix} A & \bar{D}^{(0)} \end{bmatrix} = \operatorname{im} \begin{bmatrix} T_0^{(1)} \\ J^{(1)} \end{bmatrix}$ und $\operatorname{rank} \begin{bmatrix} T_0^{(1)} \\ J^{(1)} \end{bmatrix} = n - m + q$, woraus wegen Lemma 3.15 (ii) die Regularität von $B_1 = B(\bar{D}^{(0)}, T_0^{(1)}, J^{(1)})$ folgt. Damit ist $\begin{bmatrix} \tilde{D}^{(1)} \\ \tilde{K}^{(1)} \end{bmatrix}$ definiert und wegen der Spaltenregularität von $T_0^{(1)}$ auch $\tilde{D}^{(1)}$ spaltenregulär. Es gilt ebenfalls $A^T \bar{D}^{(1)} + T_0^{(1)} K^{(1)} = 0$, was wieder $\ker \begin{bmatrix} A^T & T_0^{(1)} \end{bmatrix} = \operatorname{im} \begin{bmatrix} \bar{D}^{(1)} \\ K^{(1)} \end{bmatrix}$ bedeutet. Deshalb ist wegen Lemma 3.15 (iii) auch $B_2 = B(\bar{D}^{(1)}, T_0^{(1)}, (K^{(1)})^T)$ regulär. Die Wohldefiniertheit des Algorithmus ergibt sich damit per Induktion, wobei (3.34) aus (3.24) und (3.25) folgt.

Im Fall $\operatorname{rank}(A) = m - l$ ergibt sich die Gestalt von $T_0^{(1)}$ und $\bar{D}^{(1)}$ aus (3.23) und Algorithmus 3.27, wobei die Orthogonalität von $R_1^{(1)} \left(Z_4^{(0)} \right)^T$ und $R_2^{(1)} \left(X_4^{(0)} \right)^T$ aus der

Orthonormalität der Spalten von $T_0^{(1)}$ und $\bar{D}^{(1)}$ folgt. Aus (3.23) kann dann per Induktion gefolgert werden, daß auch $\tilde{T}_0^{(i+1)}$ bzw. $\tilde{D}^{(i+1)}$ und somit auch $T_0^{(i+1)}$ bzw. $\bar{D}^{(i+1)}$ die gleiche Gestalt besitzen.

Die übrigen Aussagen folgen aus Lemma 3.20. □

Bemerkung 3.30.

(i) Aus $A\,T_0^{(i+1)} + \bar{D}^{(i)} J^{(i+1)} = 0$ und der Spaltenorthonormalität von $\bar{D}^{(i)}$ folgt $J^{(i+1)} = -\left(\bar{D}^{(i)}\right)^T A\,T_0^{(i+1)}$ und $\left(J^{(i+1)}\right)^T J^{(i+1)} = \left(T_0^{(i+1)}\right)^T A^T A\,T_0^{(i+1)}$ und aus $A^T \bar{D}^{(i+1)} + T_0^{(i+1)} K^{(i+1)} = 0$ und der Spaltenorthonormalität von $T_0^{(i+1)}$ die Beziehungen $K^{(i+1)} = -\left(T_0^{(i+1)}\right)^T A^T \bar{D}^{(i+1)}$ und $\left(K^{(i+1)}\right)^T K^{(i+1)} = \left(\bar{D}^{(i+1)}\right)^T A\,A^T \bar{D}^{(i+1)}$. Im Fall $\operatorname{rank}(A) = m = n + q - l$ ist außerdem

$$\begin{aligned} I_{n-m+q} &= -\left(T_0^{(i+1)}\right)^T A^{-1} \bar{D}^{(i)} J^{(i+1)} \\ &= \left(J^{(i+1)}\right)^T \left(\bar{D}^{(i)}\right)^T A^{-T} A^{-1} \bar{D}^{(i)} J^{(i+1)} \qquad \text{und} \\ I_l &= -\left(\bar{D}^{(i+1)}\right)^T A^{-T} T_0^{(i+1)} K^{(i+1)} \\ &= \left(K^{(i+1)}\right)^T \left(T_0^{(i+1)}\right)^T A^{-1} A^{-T} T_0^{(i+1)} K^{(i+1)}\,. \end{aligned}$$

(ii) Wie im Algorithmus 3 aus [SS97b] sind auch hier die Bedingungen (3.32) an die Starträänderungen im Algorithmus 3.27 generisch erfüllt.

(iii) Aus (3.34) ergibt sich

$$\begin{aligned} \tan\zeta_{i+2} &\le \kappa^2 \tan\zeta_{i+1} \le \kappa^{2i+3} \tan\xi_0, \\ \tan\xi_{i+1} &\le \kappa^2 \tan\xi_i \quad\ \le \kappa^{2i+2} \tan\xi_0 \end{aligned} \tag{3.35}$$

für $i = 0, 1, \ldots$. Daraus ergibt sich eine Q-lineare Konvergenz mit dem Faktor κ^2 für jede Folge $\{T_0^{(i+1)}\}$ und $\{\bar{D}^{(i+1)}\}$.

3.2.3 Konvergenz der Singulärwerte

Nachdem die Konvergenz der Winkel zwischen den Teilräumen untersucht wurde, soll nun auch untersucht werden, inwiefern die Singulärwerte von $K^{(i)}$ und $J^{(i)}$ gegen die Singulärwerte von Σ_4 konvergieren. Dabei werden zwei von der Ungleichung (35) aus [SS97b] abweichende Ungleichungen nachgewiesen und zwei weitere erwähnt. Diese führen zu vergleichbaren Konvergenzraten wie in [SS97b, (49)], jedoch zu schlechteren Abschätzungen beim Grenzübergang.

Lemma 3.31.
Es seien $0 < l < m$, $\operatorname{rank}(A) > m - l$, $\kappa = \sigma_{m-l+1}/\sigma_{m-l} < 1$, $T_0^T T_0 = I_{n-m+q}$, $\bar{D}^T \bar{D} = I_l$ *und* Z_4 *regulär. Dann gilt*

$$\begin{aligned} \frac{\left|\,\|\bar{D}^T A\,T_0\| - \|X_4^T \Sigma_4\, Z_4^{-T}\|\,\right|}{\|\Sigma_4\|} &\le \frac{\|\bar{D}^T A\,T_0 - X_4^T \Sigma_4\, Z_4^{-T}\|}{\|\Sigma_4\|} \\ &\le \frac{\|X_0^T \Sigma_0\, Z_0\|}{\|X_0\|\,\|Z_0\|\,\sigma_{m-l+1}} \sin\xi\,\sin\zeta + \frac{\sin^2\zeta}{\cos\zeta}\,. \end{aligned} \tag{3.36}$$

Bemerkung 3.32.

(i) Aus $\operatorname{rank}(A) > m - l$ folgt $0 < \kappa$ und $\|\Sigma_4\| = \sigma_{m-l+1} > 0$.

(ii) $\bar{D}^T A\, T_0 = X_0^T \Sigma_0\, Z_0 + X_4^T \Sigma_4\, Z_4$

(iii) Die Ungleichung $\kappa < 1$ ergibt sich aus der allgemeinen Voraussetzung $\sigma_{m-l+1} < \sigma_{m-l}$. Dies wird im Beweis jedoch nicht benötigt.

Beweis. Aus der Dreiecksungleichung folgt

$$\begin{aligned}\left|\|\bar{D}^T A\, T_0\| - \|X_4^T \Sigma_4\, Z_4^{-T}\|\right| &\le \|X_0^T \Sigma_0\, Z_0 + X_4^T \Sigma_4\, Z_4 - X_4^T \Sigma_4\, Z_4^{-T}\| \\ &\le \|X_0^T \Sigma_0\, Z_0\| + \|X_4^T \Sigma_4\, Z_4 - X_4^T \Sigma_4\, Z_4^{-T}\| \\ &= \|X_0^T \Sigma_0\, Z_0\| + \|X_4^T \Sigma_4\, (V_4^T T_0\, T_0^T V_4 (T_0^T V_4)^{-1} - (T_0^T V_4)^{-1})\| \\ &= \|X_0^T \Sigma_0\, Z_0\| + \|X_4^T \Sigma_4\, (V_4^T T_0\, T_0^T V_4 - I_{n-m+q})(T_0^T V_4)^{-1})\|\,.\end{aligned}$$

Unter Verwendung eines Tricks aus [LST98, Beweis des Lemmas 3.3., vergleiche auch (13) und (14)] erhält man wegen $T_0^T T_0 = I_{n-m+q}$ und $\bar{D}^T \bar{D} = I_l$

$$\begin{aligned}\|X_4^T \Sigma_4\, (V_4^T T_0\, T_0^T V_4 - I_{n-m+q})(T_0^T V_4)^{-1})\| &\le \|X_4\|\, \|\Sigma_4\|\, \|V_4^T (T_0\, T_0^T - I_{n+q-l})\, V_4 (T_0^T V_4)^{-1}\| \\ &\le \sigma_{m-l+1} \frac{\sin^2 \zeta}{\cos \zeta}\,,\end{aligned}$$

woraus unter Beachtung von $\|X_0\| = \sin \xi$ und $\|Z_0\| = \sin \zeta$ die Ungleichung (3.36) folgt. □

Lemma 3.33.
Es seien $0 < l < m$, $\operatorname{rank}(A) > m - l$ *und* $\kappa = \sigma_{m-l+1}/\sigma_{m-l} < 1$. *Dann gilt*

$$\frac{\|A\, T_0\| - \|\Sigma_4\, Z_4\|}{\|\Sigma_4\, Z_4\|} \le \frac{1}{2} \left(\frac{\|\Sigma_0\, Z_0\|}{\|\Sigma_4\, Z_4\|\, \|Z_0\|} \right)^2 \sin^2 \zeta\,.$$

Bemerkung 3.34. Wegen $T_0^T A^T A\, T_0 = Z_0^T \Sigma_0^T \Sigma_0\, Z_0 + Z_4^T \Sigma_4^T \Sigma_4\, Z_4$ ist offensichtlich $\|A\, T_0\| \ge \|\Sigma_4\, Z_4\|$ und wegen der Dreiecksungleichung und [GVL89, Theorem 2.3.1 und (2.5.8)]

$$\begin{aligned}\|A\, T_0\|^2 - \|\Sigma_4\, Z_4\|^2 &= \|T_0^T A^T A\, T_0\| - \|Z_4^T \Sigma_4^T \Sigma_4\, Z_4\| \\ &\le \|Z_0^T \Sigma_0^T \Sigma_0\, Z_0\| \\ &= \|\Sigma_0\, Z_0\|^2.\end{aligned}$$

Beweis.

$$\begin{aligned}\frac{\|A\, T_0\| - \|\Sigma_4\, Z_4\|}{\|\Sigma_4\, Z_4\|} &= \frac{\|A\, T_0\|^2 - \|\Sigma_4\, Z_4\|^2}{\|\Sigma_4\, Z_4\|\, (\|A\, T_0\| + \|\Sigma_4\, Z_4\|)} \\ &\le \frac{\|\Sigma_0\, Z_0\|^2}{2 * \|\Sigma_4\, Z_4\|^2} \\ &= \frac{1}{2} \left(\frac{\|\Sigma_0\, Z_0\|}{\|\Sigma_4\, Z_4\|\, \|Z_0\|} \right)^2 \sin^2 \zeta\end{aligned}$$

□

Bemerkung 3.35. Analog läßt sich zeigen, daß unter den entsprechenden Bedingungen

$$\begin{aligned}\frac{\left|\|\bar{D}^T A\, T_0\| - \|X_4^{-1}\Sigma_4\, Z_4\|\right|}{\|\Sigma_4\|} &\le \frac{\|\bar{D}^T A\, T_0 - X_4^{-1}\Sigma_4\, Z_4\|}{\|\Sigma_4\|} \\ &\le \frac{\|X_0^T \Sigma_0\, Z_0\|}{\|X_0\|\,\|Z_0\|\,\sigma_{m-l+1}} \sin\xi\, \sin\zeta + \frac{\sin^2\xi}{\cos\xi}\end{aligned}$$

und

$$\frac{\|A^T\,\bar{D}\| - \|\Sigma_4^T X_4\|}{\|\Sigma_4^T X_4\|} \le \frac{1}{2}\left(\frac{\|\Sigma_0^T X_0\|}{\|\Sigma_4^T X_4\|\,\|X_0\|}\right)^2 \sin^2\xi$$

gilt.

Wenn dies auf $J^{(i+1)}$ und $K^{(i+1)}$ angewendet wird, ergibt sich zusammen mit Bemerkung 3.30

Lemma 3.36.
Es seien $0 < l < m$, $\operatorname{rank}(A) \ge m - l$, $\kappa = \sigma_{m-l+1}/\sigma_{m-l} < 1$, *(3.32) sei erfüllt und* $T_0^{(i+1)}$, $J^{(i+1)}$, $\bar{D}^{(i+1)}$ *und* $K^{(i+1)}$ *seien für* $i = 0, 1, \ldots$ *die Iterierten aus Algorithmus 3.27. Dann gilt für* $\operatorname{rank}(A) = m - l$

$$\pm\Sigma_4 = J^{(i+1)} = (K^{(i+1)})^T = 0 \tag{3.37}$$

und für $\operatorname{rank}(A) > m - l$

(3.38)

$$\begin{aligned}\frac{\left|\|K^{(i+1)}\| - \left\|\left(X_4^{(i+1)}\right)^T \Sigma_4 \left(Z_4^{(i+1)}\right)^{-T}\right\|\right|}{\|\Sigma_4\|} &\le \frac{\left\|\left(K^{(i+1)}\right)^T + \left(X_4^{(i+1)}\right)^T \Sigma_4 \left(Z_4^{(i+1)}\right)^{-T}\right\|}{\|\Sigma_4\|} \\ &\le \left(\frac{\left\|\left(X_0^{(i+1)}\right)^T \Sigma_0\, Z_0^{(i+1)}\right\|\,\kappa}{\left\|X_0^{(i+1)}\right\|\,\left\|Z_0^{(i+1)}\right\|\,\sigma_{m-l+1}} + 1\right)\kappa^{4i+2}\tan^2\xi_0 \\ \frac{\left|\|J^{(i+1)}\| - \left\|\left(X_4^{(i)}\right)^{-1}\Sigma_4\, Z_4^{(i+1)}\right\|\right|}{\|\Sigma_4\|} &\le \frac{\left\|J^{(i+1)} + \left(X_4^{(i)}\right)^{-1}\Sigma_4\, Z_4^{(i+1)}\right\|}{\|\Sigma_4\|} \\ &\le \left(\frac{\left\|\left(X_0^{(i)}\right)^T \Sigma_0\, Z_0^{(i+1)}\right\|\,\kappa}{\left\|X_0^{(i)}\right\|\,\left\|Z_0^{(i+1)}\right\|\,\sigma_{m-l+1}} + 1\right)\kappa^{4i}\tan^2\xi_0 \\ \frac{\|J^{(i+1)}\| - \left\|\Sigma_4\, Z_4^{(i+1)}\right\|}{\left\|\Sigma_4\, Z_4^{(i+1)}\right\|} &\le \frac{1}{2}\left(\frac{\left\|\Sigma_0\, Z_0^{(i+1)}\right\|}{\left\|\Sigma_4\, Z_4^{(i+1)}\right\|\,\left\|Z_0^{(i+1)}\right\|}\right)^2 \kappa^{4i+2}\tan^2\xi_0 \\ \frac{\|K^{(i+1)}\| - \left\|\Sigma_4^T X_4^{(i+1)}\right\|}{\left\|\Sigma_4^T X_4^{(i+1)}\right\|} &\le \frac{1}{2}\left(\frac{\left\|\Sigma_0^T X_0^{(i+1)}\right\|}{\left\|\Sigma_4^T X_4^{(i+1)}\right\|\,\left\|X_0^{(i+1)}\right\|}\right)^2 \kappa^{4i+4}\tan^2\xi_0\,.\end{aligned}$$

Beweis. Aus Satz 3.29 folgt (3.37).

Es sei nun $\operatorname{rank}(A) > m - l$. Wegen (3.32) und Lemma 3.20 sind $Z_4^{(i+1)}$ und $X_4^{(i)}$ regulär, da durch die QR-Zerlegungen in den Schritten S1.2 und S2.2 des Algorithmus 3.27 diese Regularität nicht beeinflußt wird. Dann sind die Voraussetzungen von Lemma 3.31, Lemma 3.33 und Bemerkung 3.35 erfüllt. Offensichtlich sind für ξ und ζ die Ungleichungen $\sin\xi \leq \tan\xi$ und $\sin\zeta \leq \tan\zeta$ erfüllt. Daraus folgen mit Bemerkung 3.30 die Ungleichungen (3.38). □

Bemerkung 3.37. Es gilt

$$\cos\xi_i = \left\| \left(X_4^{(i)} \right)^{-1} \right\|^{-1} \leq \left\| X_4^{(i)} \right\| \leq 1 \quad \text{und analog} \quad \cos\zeta_i = \left\| \left(Z_4^{(i)} \right)^{-1} \right\|^{-1} \leq \left\| Z_4^{(i)} \right\| \leq 1 \,.$$

Wegen (3.35) konvergieren somit die Singulärwerte von $X_4^{(i)}$ und $Z_4^{(i)}$ gegen Eins. Für genügend große i unterscheiden sie sich somit nur wenig von orthogonalen Matrizen. Dies gilt auch für die Transponierten, die Inversen und die Transponierten der Inversen. Die Eigenwerte, vergleiche [GVL89, P 7.1.6, S. 340], und damit auch die Singulärwerte einer Matrix hängen stetig von ihren Elementen ab. Somit unterscheiden sich für genügend große i die Singulärwerte von $\left(X_4^{(i+1)} \right)^T \Sigma_4 \left(Z_4^{(i+1)} \right)^{-T}$, $\left(X_4^{(i)} \right)^{-1} \Sigma_4 \, Z_4^{(i+1)}$, $\Sigma_4 \, Z_4^{(i+1)}$ und $\Sigma_4^T X_4^{(i+1)}$ nur beliebig wenig von den Singulärwerten von Σ_4. Damit folgt aus Lemma 3.36 die Konvergenz der Singulärwerte von $J^{(i+1)}$ bzw. $K^{(i+1)}$ gegen die Singulärwerte von Σ_4.

Daraus können folgende Konvergenzaussagen geschlossen werden.

Proposition 3.38.
Es seien die Voraussetzungen aus Lemma 3.36 und $\operatorname{rank}(A) > m - l$ *erfüllt. Dann gelten folgende Abschätzungen:*

(3.39)

$$\begin{aligned}
\frac{\left| \left\|K^{(i+1)}\right\| - \left\| \left(X_4^{(i+1)}\right)^T \Sigma_4 \left(Z_4^{(i+1)}\right)^{-T} \right\| \right|}{\|\Sigma_4\|} &\leq \frac{\left\| \left(K^{(i+1)}\right)^T + \left(X_4^{(i+1)}\right)^T \Sigma_4 \left(Z_4^{(i+1)}\right)^{-T} \right\|}{\|\Sigma_4\|} \\
&=: \rho_{2i+2} = \mathcal{O}(\kappa^{4i+2}) \\
\lim_{i\to\infty} \frac{\rho_{2i+2}}{\kappa^{4i+2}} &\leq \left(\frac{\sigma_1}{\sigma_{m-l}} + 1\right) \tan^2 \xi_0 \\
\frac{\left| \left\|J^{(i+1)}\right\| - \left\| \left(X_4^{(i)}\right)^{-1} \Sigma_4 Z_4^{(i+1)} \right\| \right|}{\|\Sigma_4\|} &\leq \frac{\left\| J^{(i+1)} + \left(X_4^{(i)}\right)^{-1} \Sigma_4 Z_4^{(i+1)} \right\|}{\|\Sigma_4\|} \\
&=: \rho_{2i+1} = \mathcal{O}(\kappa^{4i}) \\
\lim_{i\to\infty} \frac{\rho_{2i+1}}{\kappa^{4i}} &\leq \left(\frac{\sigma_1}{\sigma_{m-l}} + 1\right) \tan^2 \xi_0 \\
\frac{\left\|J^{(i+1)}\right\| - \left\|\Sigma_4 Z_4^{(i+1)}\right\|}{\left\|\Sigma_4 Z_4^{(i+1)}\right\|} &=: \tilde{\rho}_{2i+1} = \mathcal{O}(\kappa^{4i}) \\
\lim_{i\to\infty} \frac{\tilde{\rho}_{2i+1}}{\kappa^{4i}} &\leq \frac{1}{2} \left(\frac{\sigma_1}{\sigma_{m-l}}\right)^2 \tan^2 \xi_0 \\
\frac{\left\|K^{(i+1)}\right\| - \left\|\Sigma_4^T X_4^{(i+1)}\right\|}{\left\|\Sigma_4^T X_4^{(i+1)}\right\|} &=: \tilde{\rho}_{2i+2} = \mathcal{O}(\kappa^{4i+2}) \\
\lim_{i\to\infty} \frac{\tilde{\rho}_{2i+2}}{\kappa^{4i+2}} &\leq \frac{1}{2} \left(\frac{\sigma_1}{\sigma_{m-l}}\right)^2 \tan^2 \xi_0
\end{aligned}$$

Beweis. Offensichtlich gilt

$$\begin{aligned}
\left\| \left(X_0^{(i+1)}\right)^T \Sigma_0 Z_0^{(i+1)} \right\| &\leq \left\|X_0^{(i+1)}\right\| \|\Sigma_0\| \left\|Z_0^{(i+1)}\right\| \leq \left\|X_0^{(i+1)}\right\| \sigma_1 \left\|Z_0^{(i+1)}\right\| , \\
\left\| \left(X_0^{(i)}\right)^T \Sigma_0 Z_0^{(i+1)} \right\| &\leq \left\|X_0^{(i)}\right\| \|\Sigma_0\| \left\|Z_0^{(i+1)}\right\| \leq \left\|X_0^{(i)}\right\| \sigma_1 \left\|Z_0^{(i+1)}\right\| , \\
\left\|\Sigma_0 Z_0^{(i+1)}\right\| &\leq \|\Sigma_0\| \left\|Z_0^{(i+1)}\right\| \leq \sigma_1 \left\|Z_0^{(i+1)}\right\| \quad \text{und} \\
\left\|\Sigma_0^T X_0^{(i+1)}\right\| &\leq \|\Sigma_0\| \left\|X_0^{(i+1)}\right\| \leq \sigma_1 \left\|X_0^{(i+1)}\right\| .
\end{aligned} \tag{3.40}$$

Wegen Bemerkung 3.37 ist außerdem

$$\lim_{i\to\infty} \left\|\Sigma_4 Z_4^{(i+1)}\right\| = \|\Sigma_4\| = \sigma_{m-l+1} \quad \text{und} \quad \lim_{i\to\infty} \left\|\Sigma_4^T X_4^{(i+1)}\right\| = \|\Sigma_4\| = \sigma_{m-l+1} .$$

Unter Beachtung von $\kappa = \sigma_{m-l+1}/\sigma_{m-l} < 1$ folgen daraus wegen (3.38), d. h. aus Lemma 3.36, die Beziehungen (3.39), d. h. die Aussage von Proposition 3.38. □

Bemerkung 3.39. Zusammen mit Lemma 3.24 und Bemerkung 3.26 können für reguläre Matrizen A und $X_3^{(0)} \neq 0$ vermutlich aus Lemma 3.36 schärfere Abschätzungen der Grenzwerte in (3.39) gefolgert werden. Konkret kann vermutlich σ_1 durch σ_{m-l} und somit σ_1/σ_{m-l} durch 1 ersetzt werden. In diesem Fall gilt nämlich:

Falls $\sigma_1 = \sigma_{m-l}$, sind

$$\frac{\left\| \left(X_0^{(i+1)}\right)^T \Sigma_0 \, Z_0^{(i+1)} \right\|}{\left\| X_0^{(i+1)} \right\| \left\| Z_0^{(i+1)} \right\|} \leq \sigma_{m-l}, \quad \frac{\left\| \left(X_0^{(i)}\right)^T \Sigma_0 \, Z_0^{(i+1)} \right\|}{\left\| X_0^{(i)} \right\| \left\| Z_0^{(i+1)} \right\|} \leq \sigma_{m-l} \qquad \text{und}$$

$$\frac{\left\| \Sigma_0 \, Z_0^{(i+1)} \right\|}{\left\| Z_0^{(i+1)} \right\|} = \frac{\left\| \Sigma_0^T X_0^{(i+1)} \right\|}{\left\| X_0^{(i+1)} \right\|} = \sigma_{m-l} \, .$$

Bei $\sigma_1 > \sigma_{m-l}$ sind wegen Satz 3.29 die Matrizen B_i für $i = 0, 1, \ldots$ regulär. Damit sind für $B = B_0$ wegen $X_3^{(0)} \neq 0$ die Voraussetzungen von Lemma 3.24 erfüllt. Damit existiert die reguläre Matrix $X_j^{(0)}$, so daß wegen (3.30), (3.31) und Algorithmus 3.27, S1.2

$$\begin{aligned} \left\| Z_1^{(1)} \left(Z_j^{(1)}\right)^{-1} \right\| &= \left\| \tilde{Z}_1^{(1)} \left(R_1^{(1)}\right)^{-1} \left(\tilde{Z}_j^{(1)} \left(R_1^{(1)}\right)^{-1} \right)^{-1} \right\| = \left\| \tilde{Z}_1^{(1)} \left(\tilde{Z}_j^{(1)}\right)^{-1} \right\| \\ &\leq \frac{\sigma_{m-l}}{\sigma_1} \left\| X_1^{(0)} \left(X_j^{(0)}\right)^{-1} \right\| \end{aligned}$$

und falls vorhanden

$$\left\| Z_2^{(1)} \left(Z_j^{(1)}\right)^{-1} \right\| = \left\| \tilde{Z}_2^{(1)} \left(\tilde{Z}_j^{(1)}\right)^{-1} \right\| \leq \frac{\sigma_{m-l}}{\min\{\sigma_j : \sigma_j > \sigma_{m-l}\}} \left\| X_2^{(0)} \left(X_j^{(0)}\right)^{-1} \right\|$$

gilt. Aus Bemerkung 3.26 folgt daraus

$$\left\| X_1^{(1)} \left(X_j^{(1)}\right)^{-1} \right\| \leq \left(\frac{\sigma_{m-l}}{\sigma_1}\right)^2 \left\| X_1^{(0)} \left(X_j^{(0)}\right)^{-1} \right\|$$

und falls vorhanden

$$\left\| X_2^{(1)} \left(X_j^{(1)}\right)^{-1} \right\| \leq \left(\frac{\sigma_{m-l}}{\min\{\sigma_j : \sigma_j > \sigma_{m-l}\}}\right)^2 \left\| X_2^{(0)} \left(X_j^{(0)}\right)^{-1} \right\| \, .$$

Daraus folgt mit Lemma 3.24 und Bemerkung 3.26 induktiv, daß die Spektralnormen der Matrizen $Z_1^{(i)} \left(Z_j^{(i)}\right)^{-1}$, $X_1^{(i)} \left(X_j^{(i)}\right)^{-1}$ und falls vorhanden der Matrizen $Z_2^{(i)} \left(Z_j^{(i)}\right)^{-1}$ und $X_2^{(1)} \left(X_j^{(1)}\right)^{-1}$ gegen Null konvergieren. Somit konvergieren diese Matrizen gegen die Nullmatrizen entsprechender Dimension.

Für $l = n - m + q = 1$ kann die Vermutung auch nachgewiesen werden. In diesem Fall besteht $X_j^{(i)}$ aus einer Komponente von $X_3^{(i)}$, $Z_j^{(i)}$ aus der entsprechenden Komponenten von $Z_3^{(i)}$, weshalb $X_3^{(i)} \left(X_j^{(i)}\right)^{-1}$ und $Z_3^{(i)} \left(Z_j^{(i)}\right)^{-1}$ mindestens eine Eins enthält. Somit ist

die Norm dieser Vektorfolgen nach unten durch Eins beschränkt. Daraus folgt

$$\begin{aligned}
\lim_{i\to\infty} \frac{\left\| \left(X_0^{(i+1)}\right)^T \Sigma_0 Z_0^{(i+1)} \right\|}{\left\|X_0^{(i+1)}\right\| \left\|Z_0^{(i+1)}\right\|} &= \lim_{i\to\infty} \frac{\left| \left(X_j^{(i+1)}\right)^{-1} \left(X_0^{(i+1)}\right)^T \Sigma_0 Z_0^{(i+1)} \left(Z_j^{(i+1)}\right)^{-1} \right|}{\left\|X_0^{(i+1)} \left(X_j^{(i+1)}\right)^{-1}\right\| \left\|Z_0^{(i+1)} \left(Z_j^{(i+1)}\right)^{-1}\right\|} \\
&\le \lim_{i\to\infty} \frac{\sum_{\iota=1}^{3} \left| \left(X_j^{(i+1)}\right)^{-1} \left(X_\iota^{(i+1)}\right)^T \Sigma_\iota Z_\iota^{(i+1)} \left(Z_j^{(i+1)}\right)^{-1} \right|}{\sqrt{\sum_{\iota=1}^{3} \left\|X_\iota^{(i+1)} \left(X_j^{(i+1)}\right)^{-1}\right\|^2} \sqrt{\sum_{\iota=1}^{3} \left\|Z_\iota^{(i+1)} \left(Z_j^{(i+1)}\right)^{-1}\right\|^2}} \\
&= \lim_{i\to\infty} \frac{\left| \left(X_j^{(i+1)}\right)^{-1} \left(X_3^{(i+1)}\right)^T \Sigma_3 Z_3^{(i+1)} \left(Z_j^{(i+1)}\right)^{-1} \right|}{\left\|X_3^{(i+1)} \left(X_j^{(i+1)}\right)^{-1}\right\| \left\|Z_3^{(i+1)} \left(Z_j^{(i+1)}\right)^{-1}\right\|} \\
&\le \|\Sigma_3\| = \sigma_{m-l}
\end{aligned}$$

Analog läßt sich

$$\lim_{i\to\infty} \frac{\left\| \left(X_0^{(i)}\right)^T \Sigma_0 Z_0^{(i+1)} \right\|}{\left\|X_0^{(i)}\right\| \left\|Z_0^{(i+1)}\right\|} \le \sigma_{m-l} \quad \text{und} \quad \lim_{i\to\infty} \frac{\left\|\Sigma_0 Z_0^{(i+1)}\right\|^2}{\left\|Z_0^{(i+1)}\right\|^2} = \lim_{i\to\infty} \frac{\left\|\Sigma_0^T X_0^{(i+1)}\right\|^2}{\left\|X_0^{(i+1)}\right\|^2} = \sigma_{m-l}^2$$

nachweisen. Somit kann in (3.40) in diesem Fall σ_1 durch σ_{m-l} und somit σ_1/σ_{m-l} durch 1 ersetzt werden. In (3.39) stehen in diesem Fall damit die Grenzwerte $2\tan^2\xi_0$, was in etwa [SS97b, (49)] entspricht, bzw. $\tan^2\xi_0/2$.

3.2.4 Ein Startschritt eines komplexeren Algorithmus

Es sei $r := \operatorname{rank}(A)$, wobei $m-l \le r \le \min\{m, n+q-l\}$ gilt. Bei $m-l < r < \min\{m, n+q-l\}$ wird im Algorithmus 3.27 im ersten Schritt in exakter Arithmetik i. allg. nicht im V_4 und im U_4 berechnet. Nach Satz 3.29 werden jedoch in jedem Schritt auch Nullraumvektoren von A und A^T berechnet. Dies ist nicht notwendig und erhöht lediglich den Rechenaufwand. Weiterhin ist es für die Übertragung von Proposition 9 aus [SS97b] auf den hier betrachteten Fall sinnvoll, daß A quadratisch ist. Somit sollten im ersten Schritt eine Basis von $\ker A$ und von $\ker A^T$ bestimmt und anschließend die mit diesen Basen geränderte Matrix A als neue Matrix ${}_{neu}A$ betrachtet werden. Um durch Ränderung von ${}_{neu}A$ eine optimale Ränderung von A zu erhalten, sollten die Singulärwerte von ${}_{neu}A$, die zu den Singulärwerten von A hinzukommen, zwischen σ_{m-l} und σ_1 liegen. Wegen Bemerkung 3.11 (iii) bzw. den Sätzen 3.7 und 3.10 ist es dafür sinnvoll, daß diese Basen orthogonal sind und die Norm dieser Basisvektoren zwischen σ_{m-l} und σ_1 liegt. Neben der Bestimmung dieser Basen sollte somit im ersten Schritt ein solcher Wert ermittelt werden.

Zur Bestimmung von ${}_{neu}A$ mit den gewünschten Eigenschaften kann folgender Algorithmus verwendet werden.

Algorithmus 3.40.

Initialisierung:

S0: Wähle ${}_{beg}\bar{D}^{(-1)} \in \mathbb{R}^{m\times(l+1)}$ so, daß die Spalten von ${}_{beg}\bar{D}^{(-1)}$ orthonormal sind und ${}_{beg}T_0^{(-1)} \in \mathbb{R}^{(n+q-l)\times(n-m+q+1)}$, ${}_{beg}K^{(-1)} \in \mathbb{R}^{(n-m+q+1)\times(l+1)}$, setze $i := -1$

Ein Schritt $-1 \to 0$:

S1.1: Löse $\begin{bmatrix} A & {}_{beg}\bar{D}^{(-1)} \\ ({}_{beg}T_0^{(-1)})^T & {}_{beg}K^{(-1)} \end{bmatrix} \begin{bmatrix} {}_{beg}\tilde{T}_0^{(0)} \\ {}_{beg}\tilde{J}^{(0)} \end{bmatrix} = \begin{bmatrix} 0 \\ I_{n-m+q+1} \end{bmatrix}$ für $\left({}_{beg}\tilde{\boldsymbol{T}}^{(0)}\right)^T := \begin{bmatrix} {}_{beg}\tilde{T}_0^{(0)} \\ {}_{beg}\tilde{J}^{(0)} \end{bmatrix} \in \mathbb{R}^{(n+q+1)\times(n-m+q+1)}$

S1.2: Berechne ${}_rT_0^{(0)} \in \mathbb{R}^{(n+q-l)\times(r-m+l+1)}$, ${}_sT_0^{(0)} \in \mathbb{R}^{(n+q-l)\times(n+q-l-r)}$ und ${}_rJ^{(0)} \in \mathbb{R}^{(l+1)\times(r-m+l+1)}$, wobei $\ker\begin{bmatrix} A & {}_{beg}\bar{D}^{(-1)} \end{bmatrix} = \operatorname{im}\begin{bmatrix} {}_rT_0^{(0)} & {}_sT_0^{(0)} \\ {}_rJ^{(0)} & 0 \end{bmatrix}$, die Spalten von $\begin{bmatrix} {}_rT_0^{(0)} & {}_sT_0^{(0)} \end{bmatrix}$ orthonormal sind und ${}_rJ^{(0)}$ eine spaltenreguläre untere Dreiecksmatrix ist, z. B. durch QR-Faktorisierung von ${}_{beg}\tilde{T}_0^{(0)} = {}_{beg}T_0^{(0)}\,{}_{beg}R_1^{(0)}$ und entsprechender Transformation ${}_{beg}J^{(0)} := {}_{beg}\tilde{J}^{(0)}\left({}_{beg}R_1^{(0)}\right)^{-1}$ analog zu Algorithmus 3.27, S1.2, anschließender QR-Zerlegung von $\left({}_{beg}J^{(0)}\right)^T$ und entsprechender Transformation von ${}_{beg}T_0^{(0)}$

S1.3: Berechne den größten Singulärwert $\check{\sigma}$ von ${}_rJ^{(0)}$, z. B. mittels Vektoriteration

S2.1: Löse $\begin{bmatrix} A^T & \check{\sigma}\,{}_sT_0^{(0)} & {}_rT_0^{(0)}N_{r-m+l+1,r-m+l} \\ N_{l+1,l}^T({}_{beg}\bar{D}^{(-1)})^T & 0 & N_{l+1,l}^T{}_rJ^{(0)}N_{r-m+l+1,r-m+l} \end{bmatrix} \begin{bmatrix} {}_{beg}\tilde{D}^{(0)} \\ {}_{beg}\tilde{K}^{(0)} \end{bmatrix} = \begin{bmatrix} 0 \\ I_l \end{bmatrix}$ für ${}_{beg}\tilde{D}^{(0)} \in \mathbb{R}^{(m+n+q-l-r)\times l}$ und ${}_{beg}\tilde{K}^{(0)} \in \mathbb{R}^{(r-m+l)\times l}$

S2.2: Berechne analog zu S1.2 ${}_r\bar{D}^{(0)} \in \mathbb{R}^{(m+n+q-l-r)\times(r-m+l)}$, ${}_s\bar{D}^{(0)} \in \mathbb{R}^{(m+n+q-l-r)\times(m-r)}$ und ${}_rK^{(0)} \in \mathbb{R}^{(r-m+l)\times(r-m+l)}$, wobei

$$\ker\begin{bmatrix} A^T & \check{\sigma}\,{}_sT_0^{(0)} & {}_rT_0^{(0)}N_{r-m+l+1,r-m+l} \end{bmatrix} = \operatorname{im}\begin{bmatrix} {}_r\bar{D}^{(0)} & {}_s\bar{D}^{(0)} \\ {}_rK^{(0)} & 0 \end{bmatrix}, \tag{3.41}$$

die Spalten von $\begin{bmatrix} {}_r\bar{D}^{(0)}\,{}_s\bar{D}^{(0)} \end{bmatrix}$ orthonormal sind und ${}_rK^{(0)}$ eine reguläre untere Dreiecksmatrix ist

Initialisierung eines zum Algorithmus 3.27 analogen Algorithmus:

S3: Setze $i := 0$, ${}_{neu}A := \begin{bmatrix} A \\ \check{\sigma}\left({}_sT_0^{(0)}\right)^T & \check{\sigma}\,{}_s\bar{D}^{(0)} \end{bmatrix} \in \mathbb{R}^{(m+n+q-l-r)\times(m+n+q-l-r)}$, ${}_{neu}\bar{D}^{(0)} := {}_r\bar{D}^{(0)}$, ${}_{neu}T_0^{(0)} := M_{m+n+q-l-r,n+q-l}\,{}_rT_0^{(0)}N_{r-m+l+1,r-m+l}$, ${}_{neu}T_0^{(0)} \in \mathbb{R}^{(m+n+q-l-r)\times(r-m+l)}$, ${}_{neu}K^{(0)} := {}_rK^{(0)}$, ${}_{neu}l := r-m+l$, ${}_{neu}m := m+n+q-l-r$, ${}_{neu}n + {}_{neu}q := n+q$

Ende

Dieser Algorithmus besitzt folgende Eigenschaften:

Satz 3.41.
Es sei $0 < l < m$, $m - l < r = \operatorname{rank}(A) < \min\{m, n+q-l\}$ *und* $\kappa = \sigma_{m-l+1}/\sigma_{m-l} < 1$. ${}_{beg}\bar{D}^{(-1)}$, ${}_{beg}T_0^{(-1)}$ *und* ${}_{beg}K^{(-1)}$ *seien so gewählt, daß* $U_4^T\,{}_{beg}\bar{D}^{(-1)}N_{l+1,l}$ *und* $B({}_{beg}\bar{D}^{(-1)}, {}_{beg}T_0^{(-1)}, ({}_{beg}K^{(-1)})^T)$ *regulär sind und* $\left({}_{beg}\bar{D}^{(-1)}\right)^T {}_{beg}\bar{D}^{(-1)} = I_{l+1}$.

Dann ist Algorithmus 3.40 wohldefiniert, d. h. insbesondere, die Matrix $\begin{bmatrix} A^T & \check{\sigma}\,{}_sT_0^{(0)} & {}_rT_0^{(0)}N_{r-m+l+1,r-m+l} \\ N_{l+1,l}^T({}_{beg}\bar{D}^{(-1)})^T & 0 & N_{l+1,lr}^T J^{(0)} N_{r-m+l+1,r-m+l} \end{bmatrix}$ *ist regulär, die Spalten von* ${}_{beg}\tilde{T}_0^{(0)}$ *und* ${}_{beg}\tilde{D}^{(0)}$ *sind linear unabhängig,* $\operatorname{rank}{}_{beg}\tilde{J}^{(0)} = r - m + l + 1$ *und* $\operatorname{rank}{}_{beg}\tilde{K}^{(0)} = r - m + l$. *Weiterhin ist* $\sigma_1 \geq \check{\sigma} \geq \sigma_{m-l}$ *und die Voraussetzungen von Satz 3.29 sind für die in S3 des Algorithmus 3.40 definierten Matrizen* ${}_{neu}A$, ${}_{neu}\bar{D}^{(0)}$, ${}_{neu}T_0^{(0)}$ *und* ${}_{neu}K^{(0)}$ *erfüllt. Zusätzlich sind die Spalten von* ${}_{neu}\bar{D}^{(0)}$ *orthonormal und* ${}_{neu}A$ *regulär.*

Beweis. Da ${}_{beg}\bar{D}^{(-1)}$ und $\left({}_{beg}\tilde{\boldsymbol{T}}^{(0)}\right)^T$ spaltenregulär sind, ist auch ${}_{beg}\tilde{T}_0^{(0)}$ spaltenregulär. Weiterhin ist $\ker\left[A \quad {}_{beg}\bar{D}^{(-1)}\right] = \operatorname{im}\left({}_{beg}\tilde{\boldsymbol{T}}^{(0)}\right)^T$ und somit $\ker A \subseteq \operatorname{im}{}_{beg}\tilde{T}_0^{(0)}$. Wegen $\left({}_{beg}\tilde{J}^{(0)}\right)^T {}_{beg}\tilde{J}^{(0)} = \left({}_{beg}\tilde{T}_0^{(0)}\right)^T A^T A\,{}_{beg}\tilde{T}_0^{(0)}$ ist somit

$$\operatorname{rank}{}_{beg}\tilde{J}^{(0)} = n - m + q + 1 - (n + q - l - r) = r - m + l + 1\,.$$

Damit existieren ${}_rT_0^{(0)}$, ${}_sT_0^{(0)}$ und ${}_rJ^{(0)}$ mit den in S1.2 angegebenen Eigenschaften, insbesondere $\ker\left[A \quad {}_{beg}\bar{D}^{(-1)}\right] = \operatorname{im}\begin{bmatrix} {}_rT_0^{(0)} & {}_sT_0^{(0)} \\ {}_rJ^{(0)} & 0 \end{bmatrix}$. Damit ist $\operatorname{im}\begin{bmatrix} \check{\sigma}\,{}_sT_0^{(0)} & {}_rT_0^{(0)}N_{r-m+l+1,r-m+l} \\ 0 & {}_rJ^{(0)}N_{r-m+l+1,r-m+l} \end{bmatrix} \subset \ker\left[A \quad {}_{beg}\bar{D}^{(-1)}\right]$, wobei $\left(e^{n+q-l+1}\right)^T \begin{bmatrix} \check{\sigma}\,{}_sT_0^{(0)} & {}_rT_0^{(0)}N_{r-m+l+1,r-m+l} \\ 0 & {}_rJ^{(0)}N_{r-m+l+1,r-m+l} \end{bmatrix} = 0_{\mathbb{R}^{n-m+q}}^T$ ist. Deshalb ist $\operatorname{im}\begin{bmatrix} \check{\sigma}\,{}_sT_0^{(0)} & {}_rT_0^{(0)}N_{r-m+l+1,r-m+l} \\ 0 & N_{l+1,lr}^T J^{(0)} N_{r-m+l+1,r-m+l} \end{bmatrix} \subseteq \ker\left[A \quad {}_{beg}\bar{D}^{(-1)}N_{l+1,l}\right]$. Dimensionsvergleich ergibt unter Beachtung der Regularität von $U_4^T\,{}_{beg}\bar{D}^{(-1)}N_{l+1,l}$ und der Spaltenregularität von $\begin{bmatrix} {}_rT_0^{(0)} & {}_sT_0^{(0)} \end{bmatrix}$

$$\ker\left[A \quad {}_{beg}\bar{D}^{(-1)}N_{l+1,l}\right] = \operatorname{im}\begin{bmatrix} \check{\sigma}\,{}_sT_0^{(0)} & {}_rT_0^{(0)}N_{r-m+l+1,r-m+l} \\ 0 & N_{l+1,lr}^T J^{(0)} N_{r-m+l+1,r-m+l} \end{bmatrix} \tag{3.42}$$

und wegen Lemma 3.15 (ii)

(3.43) die Regularität von $\begin{bmatrix} A^T & \check{\sigma}\,{}_sT_0^{(0)} & {}_rT_0^{(0)}N_{r-m+l+1,r-m+l} \\ N_{l+1,l}^T({}_{beg}\bar{D}^{(-1)})^T & 0 & N_{l+1,lr}^T J^{(0)} N_{r-m+l+1,r-m+l} \end{bmatrix}$.

Da

$$\left[\left({}_{beg}\tilde{D}^{(0)}\right)^T \quad \left({}_{beg}\tilde{K}^{(0)}\right)^T\right]^T \quad \text{und} \quad {}_rT_0^{(0)}N_{r-m+l+1,r-m+l} \quad \text{spaltenregulär sind,} \tag{3.44}$$

ist auch ${}_{beg}\tilde{D}^{(0)}$ wegen $\left[A^T \quad \check{\sigma}\,{}_sT_0^{(0)}\right]{}_{beg}\tilde{D}^{(0)} + {}_rT_0^{(0)}N_{r-m+l+1,r-m+l}\,{}_{beg}\tilde{K}^{(0)} = 0$ spaltenregulär. Dimensionsvergleich ergibt wieder unter Beachtung von (3.43) und (3.44) $\ker\left[A^T \quad \check{\sigma}\,{}_sT_0^{(0)} \quad {}_rT_0^{(0)}N_{r-m+l+1,r-m+l}\right] = \operatorname{im}\left[\left({}_{beg}\tilde{D}^{(0)}\right)^T \quad \left({}_{beg}\tilde{K}^{(0)}\right)^T\right]^T$ und zusammen mit Lemma 3.20 (ii) (b) $\operatorname{rank}{}_{beg}\tilde{K}^{(0)} = l - (m-r) = r - m + l$. Damit existieren ${}_r\bar{D}^{(0)}$, ${}_s\bar{D}^{(0)}$ und ${}_rK^{(0)}$ mit den in S2.2 angegebenen Eigenschaften.

$$\check{\sigma} = \|{}_rJ^{(0)}\| = \|A\,{}_rT_0^{(0)}\| = \max_{\substack{z\,\in\,\operatorname{im}\,{}_rT_0^{(0)}\\ \|z\|=1}} \|A\,z\| \le \|A\| = \sigma_1$$

und unter Beachtung von $\operatorname{im}\,{}_rT_0^{(0)} \subseteq (\ker A)^\perp$ und [Bjö96, Theorem 1.2.6, S. 14]

$$\check{\sigma} = \max_{\substack{z\,\in\,\operatorname{im}\,{}_rT_0^{(0)}\\ \|z\|=1}} \|A\,z\| \ge \min_{\substack{\dim(H)\,=\,r-m+l+1\\ H\,\subseteq\,(\ker A)^\perp}} \max_{\substack{z\,\in\,H\\ \|z\|=1}} \|A\,z\| = \sigma_{r-(r-m+l+1)+1} = \sigma_{m-l}\,. \tag{3.45}$$

Weiterhin ist wegen $m - l < r < \min\{m, n+q-l\}$ und $l < m$ auch

$$0 < {}_{neu}l = r - m + l < r < m < m + n + q - l - r = {}_{neu}m\,. \tag{3.46}$$

Da $\ker A = \operatorname{im}\,{}_sT_0^{(0)}$ und $\ker\left[A^T \quad \check{\sigma}\,{}_sT_0^{(0)}\right] = \operatorname{im}\,{}_s\bar{D}^{(0)}$ sind die Singulärwerte von ${}_{neu}A$ gerade $\sigma_1, \sigma_2, \ldots, \sigma_r > 0$ und $(n+q-l-r+m-r = m+n+q-l-r-r)$ Singulärwerte $\check{\sigma} \ge \sigma_{m-l} > 0$ und somit $\operatorname{rank}({}_{neu}A) = m + n + q - l - r = {}_{neu}m \ge {}_{neu}m - {}_{neu}l$.

$$\begin{aligned} {}_{neu}\kappa &= \sigma_{{}_{neu}m-{}_{neu}l+1}({}_{neu}A)/\sigma_{{}_{neu}m-{}_{neu}l}({}_{neu}A)\\ &= \sigma_{m+n+q-l-r-(r-m+l)+1}({}_{neu}A)/\sigma_{m-l+m+n+q-l-r-r}({}_{neu}A)\\ &= \sigma_{m-l+1}(A)/\sigma_{m-l}(A) = \kappa\\ &< 1\,. \end{aligned}$$

Wenn u^i der Linkssingulärvektor zum Singulärwert σ_i ist, ist damit

$${}_{neu}U_4 = M_{m+n+q-l-r,m}\,\left[u^{m-l+1} \quad \cdots \quad u^r\right] = M_{m+n+q-l-r,m}\,U_4\,M_{l,r-m+l}\,.$$

Aus der Regularität von $U_4^T\,{}_{beg}\bar{D}^{(-1)}N_{l+1,l}$ folgt wegen (3.42) mit Lemma 3.20 (ii) (a) und Bemerkung 3.22 (i) die Regularität von $V_4^T\left[\check{\sigma}\,{}_sT_0^{(0)} \quad {}_rT_0^{(0)}N_{r-m+l+1,r-m+l}\right]$. Wegen (3.41) ergibt sich daraus mit Lemma 3.20 (ii) (b) und Bemerkung 3.22 (i) die Regularität von $U_4^TM_{m+n+q-l-r,m}^T\left[{}_r\bar{D}^{(0)} \quad {}_s\bar{D}^{(0)}\right]$. Da $\ker A^T = \operatorname{im} M_{m+n+q-l-r,m}^T\,{}_s\bar{D}^{(0)}$, ist ${}_{neu}U_4{}^T\,{}_s\bar{D}^{(0)} = 0$, woraus die Regularität von ${}_{neu}U_4{}^T\,{}_r\bar{D}^{(0)}$ folgt. Ebenfalls wegen (3.41) folgt mit Lemma 3.15 (iii) die Regularität von

$$\begin{bmatrix} {}_{neu}A & {}_{neu}\bar{D}^{(0)}\\ ({}_{neu}T_0^{(0)})^T & {}_{neu}K^{(0)} \end{bmatrix} = \begin{bmatrix} A & & \\ \check{\sigma}\left({}_sT_0^{(0)}\right)^T & \check{\sigma}\,{}_s\bar{D}^{(0)} & {}_r\bar{D}^{(0)}\\ N_{r-m+l+1,r-m+l}^T\left({}_rT_0^{(0)}\right)^T & 0 & {}_rK^{(0)} \end{bmatrix}.$$

Die Spalten von ${}_{neu}\bar{D}^{(0)} = {}_{r}\bar{D}^{(0)}$ sind laut Konstruktion orthonormal. Schließlich ist der kleinste Singulärwert der Matrix ${}_{neu}A \in \mathbb{R}^{(m+n+q-l-r)\times(m+n+q-l-r)}$ gerade $\sigma_r > 0$, weshalb ${}_{neu}A$ regulär ist. □

Bemerkung 3.42.

(i) Aus Lemma 3.20 (ii) (a) folgt bereits für spaltenreguläre Matrizen ${}_{beg}\bar{D}^{(-1)}$, daß $\operatorname{rank}{}_{beg}\tilde{J}^{(0)} \leq r - m + l + 1$ gilt. Die Gleichheit kann aus $\ker\left[A \;\; {}_{beg}\bar{D}^{(-1)}\right] = \operatorname{im}\left({}_{beg}\tilde{\boldsymbol{T}}^{(0)}\right)^T$ gefolgert werden. Somit ist $\left({}_{beg}\bar{D}^{(-1)}\right)^T {}_{beg}\bar{D}^{(-1)} = I_{l+1}$ für den Beweis dieser Aussage nicht erforderlich. Für $\|{}_{r}J^{(0)}\| = \|A\,{}_{r}T_0^{(0)}\|$, vergleiche Bemerkung 3.30, wird dies jedoch verwendet.

(ii) Im Fall $l = 0$ sollte der Algorithmus aus Bemerkung 3.28 angewendet werden.

(iii) Im Fall $l = m$ braucht $\check{\sigma}$ nicht berechnet zu werden. An den entsprechenden Stellen kann ein Wert verwendet werden, der nicht kleiner als $\|A\|$ und positiv ist. Deshalb kann mit der orthogonalen Matrix ${}_{beg}\bar{D}^{(-1)} \in \mathbb{R}^{m\times l}$, mit ${}_{beg}T_0^{(-1)} \in \mathbb{R}^{(n+q-l)\times(n-m+q)}$ und ${}_{beg}K^{(-1)} \in \mathbb{R}^{(n-m+q)\times l}$ gestartet werden. In diesem Fall kann mit der entsprechenden Veränderung der Dimensionen Algorithmus 3.40 analog durchgeführt werden. Dabei fallen insbesondere die Matrizen $N_{i,j}$ weg und $\check{\sigma} \geq \|A\|$. Die Aussagen vom Satz 3.41 gelten analog.

(iv) Im Fall $r = m - l$ besteht ${}_{r}T_0^{(0)}$ aus einer Spalte, während ${}_{r}\bar{D}^{(0)}$ und ${}_{r}K^{(0)}$ nicht existieren. Wie aus Satz 3.29 und Bemerkung 3.11 (iii) hervorgeht, ist dann ${}_{neu}A$ bereits eine optimale Ränderung von A. Deshalb braucht dann keine weitere Iteration erfolgen.

(v) Im Fall $r = \min\{m, n + q - l\}$ existiert ${}_{s}T_0^{(0)}$ bzw. ${}_{s}\bar{D}^{(0)}$ nicht. Ansonsten kann Algorithmus 3.40 jedoch analog durchgeführt werden.

Damit wurde gezeigt, wie erreicht werden kann, daß ${}_{neu}A$ eine reguläre Matrix ist. Dabei erfolgte die Ränderung so, daß eine optimale Ränderung von ${}_{neu}A$ der entsprechenden Dimension auch zu einer optimalen Ränderung von A führt. Zur Vereinfachung der Schreibweise wird im Folgenden die Bezeichnung ${}_{neu}$ weggelassen, sofern dies eindeutig ist.

3.2.5 Ein komplexerer Algorithmus

Aus Satz 3.29 und(3.35) folgt, daß bei der inversen Teilraumiteration nach Algorithmus 3.27 die Bildräume von $T_0^{(i)}$ bzw. von $\bar{D}^{(i)}$ gegen die von den entsprechenden Singulärvektoren aufgespannten Räume im V_4 bzw. im U_4 konvergieren. Jedoch sichert Algorithmus 3.27 im Gegensatz zur klassischen inversen Teilraumiteration mit anschließender Spaltenorthonormalisierung, vergleiche [SRS69, Satz 4.23, S. 182ff.], nicht, daß die Spalten von $T_0^{(i)}$ bzw. $\bar{D}^{(i)}$ gegen Singulärvektoren konvergieren. Wenn sie dies täten, würde außerdem wegen Bemerkung 3.30 $J^{(i)}$ bzw. $K^{(i)}$ gegen Diagonalmatrizen konvergieren, in deren Diagonale bis auf das Vorzeichen die entsprechenden Singulärwerte stehen. Damit die Spalten von $T_0^{(i)}$ bzw. $\bar{D}^{(i)}$ auch gegen die entsprechenden Singulärvektoren konvergieren, kann ein Algorithmus folgenden Typs verwendet werden. Er ist eine Übertragung der Algorithmen für die Teilraumiteration, z. B. aus [Ste69, (1.2), S. 362]. Dabei wird berücksichtigt, daß die

Inverse einer regulären oberen Dreiecksmatrix wieder eine obere Dreiecksmatrix ist. Hier wird im Gegensatz zu jenen Stellen jedoch nicht vorausgesetzt, daß A quadratisch ist. Im Fall quadratischer Matrizen werden auch im übertragenen Algorithmus obere Dreiecksmatrizen berechnet. Bei allgemein rechteckigen Matrizen A sind diese Matrizen i. allg. keine oberen Dreiecksmatrizen mehr, sondern nur verallgemeinerte obere Dreiecksmatrizen. Wenn jedoch die Spalten von $\bar{D}^{(i)}$ und $T_0^{(i)}$ so vertauscht werden, daß die erste mit der letzten, die zweite mit der vorletzten Spalte usw. vertauscht wird, werden aus diesen verallgemeinerten oberen Dreiecksmatrizen rechteckige untere Dreiecksmatrizen. Deshalb werden hier solche rechteckige untere Dreiecksmatrizen verwendet. Dabei kann diese Vertauschung durch die Multiplikation von $\bar{D}^{(i)}$ bzw. $T_0^{(i)}$ mit E_l bzw. E_{n-m+q} realisiert werden. Diese Vertauschungsmatrizen sind folgendermaßen definiert.

Definition 3.43. $E_i \in \mathbb{R}^{i\times i}$ ist die Matrix mit lauter Einsen in der Nebendiagonale und lauter Nullen sonst, d. h., falls die Elemente von E_i mit $e_{\iota,j}(\iota, j \in \{1,\ldots,i\})$ bezeichnet werden, ist

$$e_{\iota,j} := \delta_{\iota,i-j+1} = \begin{cases} 1 & \text{für } \iota + j = i+1 \\ 0 & \text{für } \iota + j \neq i+1 \end{cases}.$$

Die Verwendung unterer statt verallgemeinerter oberer Dreiecksmatrizen entspricht der Vorgehensweise im Algorithmus 3.40. Der folgende Algorithmus 3.44 und seine konkrete Realisierung Algorithmus 3.46 sind eine komplexere Variante von Algorithmus 3.27.

Algorithmus 3.44.

Initialisierung:

S0: Wähle $\bar{D}^{(0)} \in \mathbb{R}^{m\times l}$ so, daß die Spalten von $\bar{D}^{(0)}$ orthonormal sind, setze $i := 0$

Iteration: Schritt $i \to i+1$

solange { das Abbruchkriterium nicht erfüllt ist, } **führe folgende Schritte aus**

S1: Berechne $T_0^{(i+1)} \in \mathbb{R}^{(n+q-l)\times(n-m+q)}$ und $J^{(i+1)} \in \mathbb{R}^{l\times(n-m+q)}$, so daß die Spalten von $T_0^{(i+1)}$ orthonormal sind, $J^{(i+1)}$ eine untere Dreiecksmatrix ist und $\ker \begin{bmatrix} A & \bar{D}^{(i)} \end{bmatrix} = \operatorname{im} \begin{bmatrix} T_0^{(i+1)} \\ J^{(i+1)} \end{bmatrix}$

S2: Berechne $\bar{D}^{(i+1)} \in \mathbb{R}^{m\times l}$ und $K^{(i+1)} \in \mathbb{R}^{(n-m+q)\times l}$, so daß die Spalten von $\bar{D}^{(i+1)}$ orthonormal sind, $K^{(i+1)}$ eine untere Dreiecksmatrix ist und $\ker \begin{bmatrix} A^T & T_0^{(i+1)} \end{bmatrix} = \operatorname{im} \begin{bmatrix} \bar{D}^{(i+1)} \\ K^{(i+1)} \end{bmatrix}$

S3: Setze $i := i+1$

Ende

Bemerkung 3.45. Wenn A regulär und somit $l = n - m + q \leq n + q - l = m$ ist, sind $J^{(i+1)}$ und $K^{(i+1)}$ für $i = 0, 1, \ldots$ unter Beachtung von $T_0^{(i+1)} = -A^{-1}\bar{D}^{(i)}J^{(i+1)}$, $\bar{D}^{(i+1)} = -A^{-T}T_0^{(i+1)}K^{(i+1)}$ und der Spaltenregularität von $T_0^{(i+1)}$ und $\bar{D}^{(i+1)}$ reguläre untere Dreiecksmatrizen. Somit sind $E_l\,J^{(i+1)}E_l$ und $E_l\,K^{(i+1)}E_l$ reguläre obere Dreiecksmatrizen. Deshalb ist für $j = 1, \ldots, l$ unter Beachtung von $E_l\,M_{l,j} = N_{l,j}\,E_j$

$$\ker \begin{bmatrix} A & \bar{D}^{(i)}E_l \end{bmatrix} M_{m+l,m+j} = \operatorname{im} M_{m+l,m+j}^T \begin{bmatrix} T_0^{(i+1)} \\ E_l\,J^{(i+1)} \end{bmatrix} N_{l,j} \qquad \text{und}$$

$$\ker \begin{bmatrix} A^T & T_0^{(i+1)}E_l \end{bmatrix} M_{m+l,m+j} = \operatorname{im} M_{m+l,m+j}^T \begin{bmatrix} \bar{D}^{(i+1)} \\ E_l\,K^{(i+1)} \end{bmatrix} N_{l,j}\,.$$

Somit ist wegen der Spaltenorthonormalität von $T_0^{(i+1)}$ und $\bar{D}^{(i+1)}$ in diesem Fall Algorithmus 3.44 bis auf gewisse Vorzeichen eindeutig. Dies entspricht den entsprechenden Aussagen für die QR-Zerlegung, vergleiche z. B. [GVL89, Theorem 5.2.2, S. 217].

Dieser prinzipielle Algorithmus kann auf verschiedene Weise realisiert werden. Falls A regulär, $\left(\bar{D}^{(0)}\right)^T \bar{D}^{(0)} = I_l$, $K^{(0)}$ eine reguläre untere Dreiecksmatrix und $\ker \begin{bmatrix} A^T & T_0^{(0)} \end{bmatrix} = \operatorname{im} \begin{bmatrix} \bar{D}^{(0)} \\ K^{(0)} \end{bmatrix}$ wäre, könnten im Algorithmus 3.27 die Matrizen I_{n-m+q} und I_l durch geeignete obere Dreiecksmatrizen ersetzt werden. Diese lassen sich aus der im jeweiligen Schritt S1.1 bzw. S2.1 verwendeten Matrix B_i berechnen. Dann wären $\tilde{J}^{(i+1)}$ und $\tilde{K}^{(i+1)}$ untere Dreiecksmatrizen. Wenn nun S1.2 und S2.2 entsprechend modifiziert würden, so daß $R_1^{(i+1)}$ und $R_2^{(i+1)}$ und damit auch $(R_1^{(i+1)})^{-1}$ und $(R_2^{(i+1)})^{-1}$ untere Dreiecksmatrizen wären, so wären auch $J^{(i+1)}$ und $K^{(i+1)}$ untere Dreiecksmatrizen. Die genannten Voraussetzungen sind beispielsweise erfüllt, falls die Matrizen aus dem Schritt S3 aus dem Algorithmus 3.40 verwendet werden. Die Berechnung der benötigten oberen Dreiecksmatrizen ist jedoch kein Standardverfahren. Außerdem ist die Herleitung etwas komplizierter. Deshalb soll es hier nicht verwendet werden.

Eine andere Möglichkeit ist folgender Algorithmus. Da der Fall $l = 0$ bereits in der Bemerkung 3.28 vollständig behandelt wurde, wird dabei $0 < l$ vorausgesetzt.

Algorithmus 3.46.

Initialisierung:

S0: Wähle $\bar{D}^{(0)} \in \mathbb{R}^{m\times l}$ so, daß die Spalten von $\bar{D}^{(0)}$ orthonormal sind und $T_0^{(0)} \in \mathbb{R}^{(n+q-l)\times(n-m+q)}$, $K^{(0)} \in \mathbb{R}^{(n-m+q)\times l}$, setze $i := 0$

Iteration: Schritt $i \to i+1$

solange { das Abbruchkriterium nicht erfüllt ist, } **führe folgende Schritte aus**

S1.1: Löse $\begin{bmatrix} A & \bar{D}^{(i)} \\ (T_0^{(i)})^T & K^{(i)} \end{bmatrix} \begin{bmatrix} \tilde{T}_0^{(i+1)} \\ \tilde{J}^{(i+1)} \end{bmatrix} = \begin{bmatrix} 0 \\ I_{n-m+q} \end{bmatrix}$ für $\left(\tilde{\boldsymbol{T}}^{(i+1)}\right)^T := \begin{bmatrix} \tilde{T}_0^{(i+1)} \\ \tilde{J}^{(i+1)} \end{bmatrix}$, $\tilde{\boldsymbol{T}}^{(i+1)} \in \mathbb{R}^{(n-m+q)\times(n+q)}$

S1.2: Berechne $\breve{T}_0^{(i+1)} \in \mathbb{R}^{(n+q-l)\times(n-m+q)}$, $\breve{R}_1^{(i+1)} \in \mathbb{R}^{(n-m+q)\times(n-m+q)}$ aus der QR-Zerlegung $\tilde{T}_0^{(i+1)} = \breve{T}_0^{(i+1)} \breve{R}_1^{(i+1)}$

S1.3: Setze $\breve{J}^{(i+1)} := \tilde{J}^{(i+1)} \left(\breve{R}_1^{(i+1)}\right)^{-1}$

S1.4: Berechne die untere Dreiecksmatrix $J^{(i+1)} \in \mathbb{R}^{l\times(n-m+q)}$ und die orthogonale Matrix $R_1^{(i+1)} \in \mathbb{R}^{(n-m+q)\times(n-m+q)}$ aus der QR-Zerlegung $\left(\breve{J}^{(i+1)}\right)^T = R_1^{(i+1)} \left(J^{(i+1)}\right)^T$

S1.5: Setze $T_0^{(i+1)} := \breve{T}_0^{(i+1)} R_1^{(i+1)}$

S2.1: Löse $\begin{bmatrix} A^T & T_0^{(i+1)} \\ (\bar{D}^{(i)})^T & J^{(i+1)} \end{bmatrix} \begin{bmatrix} \tilde{D}^{(i+1)} \\ \tilde{K}^{(i+1)} \end{bmatrix} = \begin{bmatrix} 0 \\ I_l \end{bmatrix}$ für $\hat{D}^{(i+1)} := \begin{bmatrix} \tilde{D}^{(i+1)} \\ \tilde{K}^{(i+1)} \end{bmatrix}$, $\hat{D}^{(i+1)} \in \mathbb{R}^{(n+q)\times l}$

S2.2: Berechne $\breve{D}^{(i+1)} \in \mathbb{R}^{m\times l}$, $\breve{R}_2^{(i+1)} \in \mathbb{R}^{l\times l}$ aus der QR-Zerlegung $\tilde{D}^{(i+1)} = \breve{D}^{(i+1)} \breve{R}_2^{(i+1)}$

S2.3: Setze $\breve{K}^{(i+1)} := \tilde{K}^{(i+1)} \left(\breve{R}_2^{(i+1)}\right)^{-1}$

S2.4: Berechne die untere Dreiecksmatrix $K^{(i+1)} \in \mathbb{R}^{(n-m+q)\times l}$ und die orthogonale Matrix $R_2^{(i+1)} \in \mathbb{R}^{l\times l}$ aus der QR-Zerlegung $\left(\breve{K}^{(i+1)}\right)^T = R_2^{(i+1)} \left(K^{(i+1)}\right)^T$

S2.5: Setze $\bar{D}^{(i+1)} := \breve{D}^{(i+1)} R_2^{(i+1)}$

S3: Setze $i := i+1$

Ende

Analog zu Satz 3.29 wird das Verhalten von Algorithmus 3.46 durch folgenden Satz 3.47 beschrieben.

Satz 3.47.
Unter den Voraussetzungen von Satz 3.29 gelten die Aussagen aus Satz 3.29 bis auf die Aussagen zur Gestalt von $T_0^{(i+1)}$ und $\bar{D}^{(i+1)}$ im Fall $\operatorname{rank}(A) = m-l$ *auch für den Algorith-*

mus 3.46. Dies gilt auch im Fall $l = m$, wobei lediglich (3.34) durch $\xi_i = \zeta_i = 0$, $i = 0, 1, \ldots$ zu ersetzen ist. In diesem Fall entspricht der Regularität von $X_4^{(0)}$ die Regularität von $\bar{D}^{(0)}$, während κ nicht existiert.

Bemerkung 3.48. Die Nichtexistenz von κ im Fall $l = m$, vergleiche Bemerkung 3.22 (ii), liegt am Nichtvorhandensein von $\sigma_{m-l} = \sigma_0$.

Beweis. Der Beweis für den Fall $l < m$ geht wie der Beweis von Satz 3.29. Dabei ist zu beachten, daß die QR-Zerlegungen in den Schritten S1.4 und S2.4 stets existieren. Die dabei berechneten Matrizen besitzen gerade die angegebenen Eigenschaften. Außerdem bleibt durch die Multiplikation mit orthogonalen Matrizen die Spaltenregularität erhalten.

Im Fall $l = m$ entspricht wegen der Regularität von $U = U_4$ der Regularität von $X_4^{(0)}$ die Regularität von $\bar{D}^{(0)}$. Daraus und aus der Zeilenregularität von $\tilde{\boldsymbol{T}}^{(1)}$ folgt wegen

$$\tilde{J}^{(1)} = -\left(\bar{D}^{(0)}\right)^{-1} A\, \tilde{T}_0^{(1)} \tag{3.47}$$

die Spaltenregularität von $\tilde{T}_0^{(1)}$. Analog zum Fall $l < m$ folgt daraus die Spaltenregularität und damit die Regularität von $T_0^{(1)}$ und die Regularität von B_1. Daraus folgt wegen

$$\tilde{K}^{(1)} = -\left(T_0^{(1)}\right)^{-1} A^T \tilde{D}^{(1)} \tag{3.48}$$

wieder die Spaltenregularität von $\tilde{D}^{(1)}$. Daraus folgt die Regularität von B_2. Die Wohldefiniertheit des Algorithmus 3.46 ergibt sich per Induktion, $\xi_i = \zeta_i = 0$ für $i = 0, 1, \ldots$ aus Bemerkung 3.22 (ii), und

$$J^{(1)} = \left(K^{(1)}\right)^T = 0 \quad \Longleftrightarrow \quad \operatorname{rank}(A) = m - l$$

aus (3.47), (3.48) und den regulären Transformationen in S1.3, S1.4, S2.3 und S2.4 und

$$\operatorname{rank}(A) = m - l \quad \Longleftrightarrow \quad J^{(i+1)} = \left(K^{(i+1)}\right)^T = 0$$

aus den entsprechenden Beziehungen für $i \neq 0$. □

3.2.6 Weitere Eigenschaften des kombinierten komplexeren Algorithmus

Das Ziel des komplexeren Algorithmus 3.44 bzw. seiner konkreten Realisierung im Algorithmus 3.46 bestand ja darin, daß die Spalten von $T_0^{(i)}$ bzw. $\bar{D}^{(i)}$ gegen Singulärvektoren und somit $J^{(i)}$ bzw. $K^{(i)}$ gegen Diagonalmatrizen konvergieren. Wenn $J^{(i)}$ und somit $K^{(i)}$ nicht quadratisch sind, würde die Konvergenz bestenfalls gegen verallgemeinerten Diagonalmatrizen erfolgen. Diese verallgemeinerten Diagonalmatrizen besitzen zusätzliche Zeilen bzw. Spalten, die aus lauter Nullen bestehen. Deshalb sollten $J^{(i)}$ und somit $K^{(i)}$ quadratisch sein. Dies bedeutet, daß auch A quadratisch ist. Damit auch keine Probleme mit den Singulärwerten Null und den zugehörigen Nullraumvektoren auftreten, wird vorausgesetzt, daß A regulär ist. All dies kann erreicht werden, wenn die Matrix ${}_{neu}A$ aus Algorithmus 3.40 verwendet wird. Dies bedeutet, daß vor dem Algorithmus 3.46 Algorithmus 3.40 durchgeführt wird. Einige Eigenschaften dieser Algorithmenkombination sollen in diesem Unterabschnitt gezeigt werden. Wegen Bemerkung 3.42 braucht dabei nur der Fall $l > m - \operatorname{rank}(A) = m - r \geq 0$ betrachtet werden. Dabei ist entsprechend dieser Bemerkung in den Fällen $l = m$ bzw. $r = \operatorname{rank}(A) = \min\{m, n + q - l\}$ Algorithmus 3.40 zu modifizieren.

Satz 3.49.
Es sei $l > m - r \geq 0$. Im Fall $l < m$ sei $\kappa = \sigma_{m-l+1}/\sigma_{m-l} < 1$. ${}_{beg}\bar{D}^{(-1)}$, ${}_{beg}T_0^{(-1)}$ und ${}_{beg}K^{(-1)}$ seien so gewählt, daß im Fall $l < m$ die Matrizen $U_4^T\,{}_{beg}\bar{D}^{(-1)}N_{l+1,l}$ und $B({}_{beg}\bar{D}^{(-1)}, {}_{beg}T_0^{(-1)}, ({}_{beg}K^{(-1)})^T)$ regulär sind und $\left({}_{beg}\bar{D}^{(-1)}\right)^T {}_{beg}\bar{D}^{(-1)} = I_{l+1}$ gilt. Im Fall $l = m$ seien $U_4^T\,{}_{beg}\bar{D}^{(-1)}$ und $B({}_{beg}\bar{D}^{(-1)}, {}_{beg}T_0^{(-1)}, ({}_{beg}K^{(-1)})^T)$ regulär und $\left({}_{beg}\bar{D}^{(-1)}\right)^T {}_{beg}\bar{D}^{(-1)} = I_l$. Weiterhin sei für jedes $j \in \{1, \ldots, {}_{neu}l - 1\}$ mit $\sigma_{{}_{neu}m-j}({}_{neu}A) > \sigma_{{}_{neu}m-j+1}({}_{neu}A)$ auch $N_{{}_{neu}l,j}^T\,{}_{neu}X_4^{(0)}N_{{}_{neu}l,j}$ regulär.

Dann gilt für die im Algorithmus 3.46 nach Ausführung von Algorithmus 3.40 unter Beachtung von Bemerkung 3.42 erzeugten Iterierten, daß die Spalten von $\bar{D}^{(i+1)}$ bzw. von $T_0^{(i+1)}$ mit $i \to \infty$ gegen Links- bzw. Rechtssingulärvektoren zu den kleinsten Singulärwerten von ${}_{neu}A$ konvergieren. Dabei bilden diese Singulärvektoren orthonormale Basen der Räume dieser Links- bzw. Rechtssingulärvektoren zu diesen kleinsten Singulärwerten. $J^{(i+1)}$ und $K^{(i+1)} \in \mathbb{R}^{{}_{neu}l \times {}_{neu}l}$ konvergieren gegen Diagonalmatrizen, wobei in der Diagonalen bis auf das Vorzeichen die zugehörigen Singulärwerten stehen. Die Werte in den Diagonalmatrizen sind nach nach ihren Beträgen geordnet. Dabei steht links oben der betragsgrößte.

Bemerkung 3.50.

(i) Daß bei der Teilraumiteration für Eigenwerte mit symmetrischen, positiv definiten Matrizen mit einfachen Eigenwerten die orthonormalisierten Spalten der $\bar{D}^{(i+1)}$ bzw. $T_0^{(i+1)}$ entsprechenden Matrizen gegen Eigenvektoren konvergieren, wurde bereits in [SRS69, Satz 4.23, S. 182ff.] gezeigt. Dort wurden lediglich zusätzlich gefordert, daß die Spalten der $\bar{D}^{(0)}$ entsprechenden Matrix Komponenten in den entsprechenden Eigenrichtungen besitzen. Dies entspricht der Forderung, daß für jedes $j \in \{1, \ldots, {}_{neu}l - 1\}$ mit $\sigma_{{}_{neu}m-j}({}_{neu}A) > \sigma_{{}_{neu}m-j+1}({}_{neu}A)$ auch $N_{{}_{neu}l,j}^T\,{}_{neu}X_4^{(0)}M_{{}_{neu}l,j}$ regulär ist. Auch die Konvergenz der im gewissen Sinne $J^{(i+1)}$ bzw. $K^{(i+1)}$ entsprechenden Matrizen gegen Diagonalmatrizen mit Eigenwerten in der Hauptdiagonalen steht in diesem Satz.

(ii) Die Forderung, daß für jedes $j \in \{1, \ldots, {}_{neu}l - 1\}$ mit $\sigma_{{}_{neu}m-j}({}_{neu}A) > \sigma_{{}_{neu}m-j+1}({}_{neu}A)$ auch $N_{{}_{neu}l,j}^T\,{}_{neu}X_4^{(0)}M_{{}_{neu}l,j}$ regulär ist, kann vermutlich weggelassen werden. Im Fall von [SRS69, Satz 4.23, S. 182ff.] würde die Weglassung der entsprechenden Forderung nach dem dort verwendeten Beweis lediglich dazu führen, daß unter gewissen Regularitätsvoraussetzungen im Grenzfall die Reihenfolge der Eigenvektoren miteinander vertauscht würde. Im dort behandelten Fall verlangt dies lediglich eine entsprechende Fallunterscheidung. Im Fall von Satz 3.49 ist dies noch exakt nachzuweisen.

(iii) Sollten in ${}_{neu}A$ die letzten ${}_{neu}l = r - m + l$ Singulärwerte einfach sein, kann die Bedingung, daß für jedes $j \in \{1, \ldots, {}_{neu}l - 1\}$ mit $\sigma_{{}_{neu}m-j}({}_{neu}A) > \sigma_{{}_{neu}m-j+1}({}_{neu}A)$ auch $N_{{}_{neu}l,j}^T\,{}_{neu}X_4^{(0)}M_{{}_{neu}l,j}$ regulär ist, bei Übertragung des entsprechenden Beweises weggelassen werden. Dazu wird in ${}_{neu}X_4^{(0)}$ in der letzten Spalte die unterste nichtverschwindende Komponente gesucht. Die entsprechende Spalte von $\bar{D}^{(i+1)}$ konvergiert dann gegen den zugehörigen Linkssingulärvektor. Die Grenzwerte der übrigen Spalten stehen senkrecht darauf. Die entsprechende Vorgehensweise wird für die vorhergehenden Spalten induktiv angewandt. Dabei werden alle Zeilen nicht berücksichtigt, die bei weiter rechts stehenden Spalten ausgewählt wurden. Dies entspricht der Orthogonalität zu bereits als Grenzwert behandelten Singulärvektoren. Zur Durchführbarkeit

dieser Vorgehensweise ist lediglich noch zu sichern, daß ${}_{neu}X_4^{(i)}$ nicht gegen eine singuläre Matrix konvergiert. Der weitere Beweis erfolgt analog zum Beweis von Satz 3.49. Dabei sind die Singulärwerte nicht notwendigerweise nach ihrem Betrag, sondern nach der gemäß dieser Bemerkung ermittelten Reihenfolge der Linkssingulärvektoren geordnet.

(iv) Da sich der Beweis auf den Algorithmus 3.46 nach Ausführung des Algorithmus 3.40 bezieht, sind alle Dimensionen und die sich darauf beziehenden Indizes die im Algorithmus 3.40 neu definierten Dimensionen. Deshalb wird auch hier die Bezeichnung ${}_{neu}$ weggelassen.

(v) Der Satz 3.49 gilt ebenfalls im Fall $\sigma_{m-l+1} = \sigma_{m-l}$. In diesem Fall wird jedoch keine vollständige Basis der zu σ_{m-l+1} gehörenden Singulärvektoren berechnet.

(vi) Der Beweis erfolgt für exakte Arithmetik. Mögliche Änderungen bei Verwendung von Zahlen mit endlicher Stellenzahl werden im Anschluß erwähnt.

(vii) Um die ganzen Probleme mit den Vorzeichen zu umgehen, kann beispielsweise gefordert werden, daß die Diagonalelemente der regulären unteren Dreiecksmatrizen $J^{(i+1)}$ und $K^{(i+1)}$, vergleiche Bemerkung 3.45, positiv sind. Dann würde die Konvergenz gegen feste Grenzwerte erfolgen. Im allgemeinen Fall erfolgt die Konvergenz gegen Häufungspunkte, die sich durch das Vorzeichen der einzelnen Singulärwerte und -vektoren unterscheiden können.

Beweis. Nach Algorithmus 3.40 ist $0 < l = n - m + q \leq n + q - l = m$ und A regulär. Wegen S1.4 und S2.4 sind $J^{(i+1)}$ und $K^{(i+1)}$ für $i = 0, 1, \ldots$ untere Dreiecksmatrizen. $X_4^{(0)}$ ist wegen (3.32) und Satz 3.41 regulär. Für jedes $j \in \{1, \ldots, l-1\}$ mit $\sigma_{m-j} > \sigma_{m-j+1}$ ist $N_{l,j}^T X_4^{(0)} N_{l,j}$ regulär. Deshalb kann Algorithmus 3.46 mit den Starträndерungen $\bar{D}^{(0)} N_{l,j}$ und j geeigneten Spalten von $T_0^{(0)}$ und den entsprechenden Zeilen von $K^{(0)} N_{l,j}$ betrachtet werden. Dann sind die Voraussetzungen von Satz 3.47 erfüllt, weshalb (3.35) gilt. Da $J^{(i+1)}$ und $K^{(i+1)}$ untere Dreiecksmatrizen sind, entspricht dies bis auf Vorzeichen, vergleiche Bemerkung 3.45, der Durchführung von Algorithmus 3.44 für das volle System und anschließender Betrachtung der entsprechenden Spalten. Daraus folgt, daß $M_{m,m-j}^T U^T \bar{D}^{(i+1)} N_{l,j}$ mit $i \to \infty$ gegen die Nullmatrix konvergiert. Wegen der Spaltenregularität von $\bar{D}^{(i+1)} N_{l,j}$ ist

$$\lim_{i \to \infty} \operatorname{im} \bar{D}^{(i+1)} N_{l,j} = \operatorname{im} U N_{m,j} = \operatorname{im} U_4 N_{l,j} .$$

Wegen (3.35) gilt dies auch für $j = l$. Wegen der Orthonormalität von $\bar{D}^{(i+1)}$ konvergieren somit die Spalten von $\bar{D}^{(i+1)}$ gegen orthonormierte Linearkombinationen von Linkssingulärvektoren zum gleichen Singulärwert, also gegen Linkssingulärvektoren. Aus Dimensionsgründen ergibt sich sogar, daß diese Linkssingulärvektoren orthonormale Basen der Räume dieser Linkssingulärvektoren zu den kleinsten Singulärwerten von A bilden. Da $J^{(i+2)}$ eine reguläre untere Dreiecksmatrix, vergleiche Bemerkung 3.45, ist und die Spalten von $T_0^{(i+2)}$ orthonormal sind, konvergieren die Spalten von $T_0^{(i+1)}$ mit $i \to \infty$ gegen Rechtssingulärvektoren, die zu den Linkssingulärvektoren gehören, gegen die die entsprechenden Spalten von $\bar{D}^{(i+1)}$ konvergieren. Wegen Bemerkung 3.30 (i) konvergieren $J^{(i+1)}$ und $K^{(i+1)}$ somit gegen Diagonalmatrizen, wobei in der Diagonalen bis auf das Vorzeichen die zugehörigen Singulärwerte stehen. Aus der Ordnung der Singulärvektoren ergeben sich die übrigen Aussagen des Satzes. □

Bemerkung 3.51.

(i) Der Beweis erfolgte für exakte Arithmetik. Bei Rechnungen mit endlicher Stellenzahl führen Rundungsfehler i.allg. dazu, daß die Spalten von $\bar{D}^{(i)}$ Komponenten in die Richtung aller Spalten von U besitzen. Insbesondere ist dann für jedes
$$j \in \{1, \ldots, l-1\} \quad \text{mit} \quad \sigma_{m-j} > \sigma_{m-j+1}$$
die Matrix $N_{l,j}^T X_4^{(i)} N_{l,j}$ regulär. Deshalb braucht in diesem Fall nicht für jedes $j \in \{1, \ldots, {}_{neu}l - 1\}$ mit $\sigma_{{}_{neu}m-j}({}_{neu}A) > \sigma_{{}_{neu}m-j+1}({}_{neu}A)$ die Regularität von $N_{{}_{neu}l,j}^T \, {}_{neu}X_4^{(0)} N_{{}_{neu}l,j}$ vorausgesetzt werden.

(ii) Die Zuordnung der Rechts- und Linkssingulärvektoren zueinander kann beispielsweise bis auf das Vorzeichen über die Singulärwertzerlegung erfolgen. Wenn die Diagonalelemente von $J^{(i+1)}$ und $K^{(i+1)}$ positiv sind, wird beispielsweise einer der beiden zusammengehörigen Singulärvektoren o. B. d. A. der Rechtssingulärvektor mit -1 multipliziert.

Nun soll noch Proposition 9 aus [SS97b] auf den hier betrachteten Fall übertragen werden. Dabei werden zuerst die im Satz 3.49 nichtbehandelten Fälle betrachtet. Anschließend wird für quadratische Matrizen gezeigt, daß für $l > 0$ die Aussagen über die Konditionszahlen analog zu denen in der Proposition 9 aus [SS97b] sind. Daraus ergibt sich die entsprechende Aussage für die im Satz 3.49 behandelten Fälle.

Bemerkung 3.52.

(i) Im Fall $l = 0$ ist bei Anwendung des Algorithmus aus Bemerkung 3.28 (i), vergleiche Bemerkung 3.42 (ii), $\operatorname{cond} B_1 = \max\{\sigma_1, 1\}/\min\{\sigma_{m-l}, 1\}$, wobei sowohl B_1 mit der Ausgangsmatrix gebildet wird, als auch die σ_i die Singulärwerte der Ausgangsmatrix A sind. Wenn A jedoch nicht mit der Matrix $\boldsymbol{T}^{(1)}$, sondern mit der Matrix $\check{\sigma}\,\boldsymbol{T}^{(1)}$ gerändert wird, wobei $\sigma_1 \geq \check{\sigma} \geq \sigma_{m-l}$ ist, gilt $\operatorname{cond} B_1 = \sigma_1/\sigma_{m-l}$. Falls allgemein mit der Matrix $t\,\boldsymbol{T}^{(1)}$ mit $t > 0$ gerändert wird, ist $\operatorname{cond} B_1 = \max\{\sigma_1, t\}/\min\{\sigma_{m-l}, t\}$. Ein Wert für $\check{\sigma}$ mit $\sigma_1 \geq \check{\sigma} \geq \sigma_{m-l}$ kann z.B. mittels $\check{\sigma} := \|A\,z\|$ mit $\|z\| = 1$ und $z \in \ker \boldsymbol{T}^{(1)} = \operatorname{im} A^T$ bzw. mittels $\check{\sigma} := \|(e^i)^T A\|$, $i \in \{1, 2, \ldots, m\}$, d. h. der Norm einer Zeile von A ermittelt werden.

(ii) Im Fall $l > 0$ und $r = m - l$ ist nach Ausführung der entsprechenden Modifikation von Algorithmus 3.40 $B_0 = {}_{neu}A$ und $\operatorname{cond} B_0 = \sigma_1/\sigma_{m-l}$, vergleiche Bemerkung 3.42 (iv).

Proposition 3.53.
Es sei $0 < l \leq m$, $m = n + q - l$, $\operatorname{rank}(A) \geq m - l$, im Fall $l < m$ sei $\kappa = \sigma_{m-l+1}/\sigma_{m-l} < 1$ und die Starträndерungen $\bar{D}^{(0)}$, $T_0^{(0)}$ und $K^{(0)}$ seien so gewählt, daß die Spalten von $\bar{D}^{(0)}$ orthonormal sind und daß (3.32) erfüllt ist.

Dann erfüllen die die Matrizen B_i gemäß (3.33) mit den Matrizen $\bar{D}^{(i)}$, $T_0^{(i)}$, $K^{(i)}$ und $J^{(i)}$ aus den Algorithmen 3.27, 3.44 bzw. 3.46

$$\operatorname{cond}(B_i) \geq \operatorname{cond}(B_{i+1}) \quad (i = 0, 1, \ldots)\,, \tag{3.49}$$

im Fall $l < m$

$$\lim_{i \to \infty} \operatorname{cond}(B_i) = \frac{\max\left\{\sigma_1, \sqrt{1 + \sigma_{m-l+1}^2}\right\}}{\min\left\{\sigma_{m-l}, \sqrt{1 + \sigma_m^2}\right\}} \tag{3.50}$$

und im Fall $l = m$

$$\lim_{i\to\infty} \operatorname{cond}(B_i) = \frac{\sqrt{1+\sigma_1^2}}{\sqrt{1+\sigma_m^2}}. \tag{3.51}$$

Beweis. Für den Nachweis von (3.49) ist

$$\operatorname{cond}\underbrace{\begin{bmatrix} A & \bar{D}^{(i)} \\ (T_0^{(i)})^T & K^{(i)} \end{bmatrix}}_{B_{2i}} \geq \operatorname{cond}\underbrace{\begin{bmatrix} A^T & T_0^{(i+1)} \\ (\bar{D}^{(i)})^T & J^{(i+1)} \end{bmatrix}}_{B_{2i+1}^T} \geq \operatorname{cond}\underbrace{\begin{bmatrix} A & \bar{D}^{(i+1)} \\ (T_0^{(i+1)})^T & K^{(i+1)} \end{bmatrix}}_{B_{2i+2}} \quad (i = 0, 1, \ldots) \tag{3.52}$$

zu zeigen. Aus der Interlacing Property folgt

$$\operatorname{cond}(B_{2i}) = \frac{\sigma_1(B_{2i})}{\sigma_{n+q}(B_{2i})} \geq \frac{\sigma_1(\begin{bmatrix} A & \bar{D}^{(i)} \end{bmatrix})}{\sigma_m(\begin{bmatrix} A & \bar{D}^{(i)} \end{bmatrix})}, \tag{3.53}$$

vergleiche (3.2). Weiterhin ist $\ker \begin{bmatrix} A & \bar{D}^{(i)} \end{bmatrix} = \operatorname{im} \left(\boldsymbol{T}^{(i+1)}\right)^T$ mit $\boldsymbol{T}^{(i+1)} = \begin{bmatrix} \left(T_0^{(i+1)}\right)^T & \left(J^{(i+1)}\right)^T \end{bmatrix}$. Dies bedeutet. daß die Singulärwerte

$$\sigma_j(B_{2i+1}) = \sigma_j\left(\begin{bmatrix} A^T & T_0^{(i+1)} \\ (\bar{D}^{(i)})^T & J^{(i+1)} \end{bmatrix}\right)$$

von B_{2i+1} gerade die Singulärwerte von $\begin{bmatrix} A & \bar{D}^{(i)} \end{bmatrix}$, ergänzt durch die Singulärwerte $\sigma_j(\boldsymbol{T}^{(i+1)}) = \sigma_j(\begin{bmatrix} (T_0^{(i+1)})^T & (J^{(i+1)})^T \end{bmatrix}) = \sqrt{1+\left[\sigma_j(J^{(i+1)})\right]^2}$ sind. Folglich ergibt sich

$$\operatorname{cond}(B_{2i+1}) = \frac{\max\{\sigma_1(\begin{bmatrix} A & \bar{D}^{(i)} \end{bmatrix}), \sigma_1(\boldsymbol{T}^{(i+1)})\}}{\min\{\sigma_m(\begin{bmatrix} A & \bar{D}^{(i)} \end{bmatrix}), \sigma_l(\boldsymbol{T}^{(i+1)})\}}. \tag{3.54}$$

Aus $\|(u^m)^T \begin{bmatrix} A & \bar{D}^{(i)} \end{bmatrix}\|^2 \leq \|(u^m)^T A\|^2 + \|u^m\|^2 = \sigma_m^2 + 1$, wegen der Orthonormalität der Spalten von $\bar{D}^{(i)}$, erhält man, da A quadratisch ist und die Spalten von $T_0^{(i+1)}$ orthonormal sind,

$$\sigma_m(\begin{bmatrix} A & \bar{D}^{(i)} \end{bmatrix}) \leq \sqrt{1+\sigma_m^2} \leq \sqrt{1+\left[\sigma_l(A\, T_0^{(i+1)})\right]^2} = \sqrt{1+\left[\sigma_l(J^{(i+1)})\right]^2}, \tag{3.55}$$

weshalb der Nenner in (3.54) $\sigma_m(\begin{bmatrix} A & \bar{D}^{(i)} \end{bmatrix})$ ist. Außerdem kann aus $A\, T_0^{(i+1)} + \bar{D}^{(i)} J^{(i+1)} = 0$ gefolgert werden, daß

$$\sigma_1(J^{(i+1)}) = \|(\bar{D}^{(i)})^T A T_0^{(i+1)}\| \leq \|A^T \bar{D}^{(i)}\|,$$

woraus

$$\begin{aligned} \left[\sigma_1(\boldsymbol{T}^{(i+1)})\right]^2 = 1 + \left[\sigma_1(J^{(i+1)})\right]^2 \leq 1 + \|A^T \bar{D}^{(i)}\|^2 &= \left\| \begin{bmatrix} A^T \\ (\bar{D}^{(i)})^T \end{bmatrix} \bar{D}^{(i)} \right\|^2 \\ \leq \left\| \begin{bmatrix} A & \bar{D}^{(i)} \end{bmatrix} \right\|^2 &= \left[\sigma_1(\begin{bmatrix} A & \bar{D}^{(i)} \end{bmatrix})\right]^2 \end{aligned}$$

folgt. Somit ist der Zähler in (3.54) $\sigma_1(\begin{bmatrix} A & \bar{D}^{(i)} \end{bmatrix})$. Damit ergibt sich

$$\operatorname{cond}(B_{2i+1}) = \sigma_1(\begin{bmatrix} A & \bar{D}^{(i)} \end{bmatrix})/\sigma_m(\begin{bmatrix} A & \bar{D}^{(i)} \end{bmatrix}),$$

woraus zusammen mit (3.53) die linke Ungleichung aus (3.52) folgt. Die rechte ergibt sich analog, wenn statt B die Matrix B^T betrachtet wird. Da die Konditionszahlen $\operatorname{cond}(B_i)$ monoton fallen, existiert ein Grenzwert und dieser Grenzwert ist wegen

$$\lim_{i\to\infty} \operatorname{cond}(B_i) = \lim_{i\to\infty} \operatorname{cond}(B_{2i+1}) = \lim_{i\to\infty} \frac{\sigma_1(\begin{bmatrix} A & \bar{D}^{(i)} \end{bmatrix})}{\sigma_m(\begin{bmatrix} A & \bar{D}^{(i)} \end{bmatrix})} = \frac{\sigma_1(\begin{bmatrix} A & U_4 \end{bmatrix})}{\sigma_m(\begin{bmatrix} A & U_4 \end{bmatrix})} \tag{3.56}$$

durch (3.50) bzw. durch (3.51) gegeben. (3.56) gilt, da wegen (3.35) $\lim_{i\to\infty} \operatorname{im} \bar{D}^{(i)} = \operatorname{im} U_4$ und die Spalten von $\bar{D}^{(i)}$ orthonormiert sind, vergleiche auch Bemerkung 3.37. □

Bemerkung 3.54.

(i) $l \le m$ ist laut Definition von l, siehe (2.3), erfüllt.

(ii) Daß A quadratisch ist, wird in (3.55) verwendet.

(iii) Dagegen wird $\kappa = \sigma_{m-l+1}/\sigma_{m-l} < 1$ und die Regularität von $X_4^{(0)} = U_4^T \bar{D}^{(0)}$ lediglich in (3.56) benutzt. Die Bedingung $\kappa < 1$ kann dabei abgeschwächt werden.

Folgerung 3.55.
Es sei $l > m - r \ge 0$. Im Fall $l < m$ sei $\kappa = \sigma_{m-l+1}/\sigma_{m-l} < 1$. ${}_{beg}\bar{D}^{(-1)}$, ${}_{beg}T_0^{(-1)}$ und ${}_{beg}K^{(-1)}$ seien so gewählt, daß im Fall $l < m$ die Matrizen $U_4^T\, {}_{beg}\bar{D}^{(-1)} N_{l+1,l}$ und $B({}_{beg}\bar{D}^{(-1)}, {}_{beg}T_0^{(-1)}, ({}_{beg}K^{(-1)})^T)$ regulär sind und $\left({}_{beg}\bar{D}^{(-1)}\right)^T {}_{beg}\bar{D}^{(-1)} = I_{l+1}$ gilt. Im Fall $l = m$ seien $U_4^T\, {}_{beg}\bar{D}^{(-1)}$ und $B({}_{beg}\bar{D}^{(-1)}, {}_{beg}T_0^{(-1)}, ({}_{beg}K^{(-1)})^T)$ regulär und $\left({}_{beg}\bar{D}^{(-1)}\right)^T {}_{beg}\bar{D}^{(-1)} = I_l$.

Wenn zuerst Algorithmus 3.40 und anschließend einer der Algorithmen 3.27, 3.44 bzw. 3.46 ausgeführt wird, dann gelten für die mit ${}_{neu}A$ definierten Matrizen B_i mit $i \ge 0$ (3.49) und im Fall $l < m$

$$\lim_{i\to\infty} \operatorname{cond}(B_i) = \frac{\max\left\{\sigma_1, \sqrt{1+\sigma_{m-l+1}^2}\right\}}{\min\left\{\sigma_{m-l}, \sqrt{1+\sigma_r^2}\right\}},$$

im Fall $l = m = n = r$

$$\lim_{i\to\infty} \operatorname{cond}(B_i) = \frac{\sqrt{1+\sigma_1^2}}{\sqrt{1+\sigma_m^2}}$$

und im Fall $l = m$ und $r < \max\{m, n+q-l\}$

$$\lim_{i\to\infty} \operatorname{cond}(B_i) = \frac{\max\left\{\check{\sigma}, \sqrt{1+\sigma_1^2}\right\}}{\min\left\{\check{\sigma}, \sqrt{1+\sigma_r^2}\right\}},$$

wobei $\check{\sigma} \ge \|A\|$ der Wert aus Bemerkung 3.42 (iii) ist. Dabei sind jeweils die Singulärwerte der Ausgangsmatriz A des Algorithmus 3.40 und die entsprechenden Indizes gemeint.

Bemerkung 3.56. Da $l \leq q \leq m$ und $r \leq \min\{m, n+q-l\}$ gilt, sind die Fallunterscheidungen in der Folgerung 3.55 vollständig.

Beweis. Im Fall $l < m$ und $r < \min\{m, n+q-l\}$ sind die Voraussetzungen von Satz 3.41 erfüllt. Deshalb sind in diesem Fall die Voraussetzungen von Satz 3.29 erfüllt, die Spalten von ${}_{neu}\bar{D}^{(0)}$ sind orthonormal und ${}_{neu}A$ ist regulär und somit quadratisch. Damit sind in diesem Fall die Voraussetzungen von Proposition 3.53 erfüllt. Im Fall $l < m$ und $r = \min\{m, n+q-l\}$ existieren lediglich ${}_sT_0^{(0)}$ bzw. ${}_s\bar{D}^{(0)}$ nicht, vergleiche Bemerkung 3.42 (v). Deshalb sind auch in diesem Fall die Voraussetzungen von Proposition 3.53 erfüllt. Nach Bemerkung 3.42 (iii) gilt dies auch für den Fall $l = m$.

Somit ist Proposition 3.53 anwendbar, woraus (3.49) folgt. ${}_{neu}l = {}_{neu}m$ gilt wegen $l \leq q \leq m$, vergleiche (2.3), genau dann, wenn $l = m = r = n$ ist, vergleiche (3.46). In diesem Fall ist ${}_{neu}A = A$, ${}_{neu}m = m$ und somit $\sigma_1({}_{neu}A) = \sigma_1$ und $\sigma_{{}_{neu}m}({}_{neu}A) = \sigma_m$. Im Fall $l = m$ und $r < \max\{m, n\}$ sind die Singulärwerte von ${}_{neu}A$ gerade die Singulärwerte σ_j mit $j = 1, \ldots, r$ von A und $m - r + n - r$ Singulärwerte $\check{\sigma} \geq \|A\| = \sigma_1$ gemäß Bemerkung 3.42 (iii). Damit ist $\sigma_1({}_{neu}A) = \check{\sigma}$, $\sigma_{{}_{neu}m-{}_{neu}l}({}_{neu}A) = \sigma_{m+n-r-r}({}_{neu}A) = \check{\sigma}$, $\sigma_{{}_{neu}m-{}_{neu}l+1}({}_{neu}A) = \sigma_1$ und $\sigma_{{}_{neu}m}({}_{neu}A) = \sigma_{m+n-r}({}_{neu}A) = \sigma_r$. Schließlich ist im Fall $l < m$, vergleiche den Beweis von Satz 3.41, $\sigma_1({}_{neu}A) = \sigma_1$, $\sigma_{{}_{neu}m-{}_{neu}l}({}_{neu}A) = \sigma_{m-l}$, $\sigma_{{}_{neu}m-{}_{neu}l+1}({}_{neu}A) = \sigma_{m-l+1}$ und $\sigma_{{}_{neu}m}({}_{neu}A) = \sigma_{m+n+q-l-r}({}_{neu}A) = \sigma_r$. Daraus folgen die übrigen Aussagen von Folgerung 3.55. □

Wie in [SS97b] fallen somit die Konditionszahlen der mit ${}_{neu}A$ definierten Matrizen monoton und die Grenzwerte besitzen eine analoge Gestalt. Proposition 9 aus [SS97b] ist ein Spezialfall von Proposition 3.53 im Spezialfall $1 = q = l < m = n$.

3.2.7 Numerische Beispiele

Analog zum Test 2 aus [SS97b] wurde die Kombination aus Algorithmus 3.40 und Algorithmus 3.46 unter Beachtung von Bemerkung 3.42 für verschiedene zufällig erzeugte Matrizen und Startränderungen getestet. Alle Berechnungen erfolgten in IEEE Arithmetik mit doppelter Genauigkeit, d. h. mit der relativen Maschinengenauigkeit von $\approx 2.2 * 10^{-16}$, mittels MATLAB. Dabei wurden für $m, n+q-l \in \{2;4;6;8;10;15;20;25;30\}$, $l \in [0; \min\{m;7\}]$ unter Beachtung von $n+q-l \geq m+1-l$, vergleiche Definition 2.1, und für alle Werte von $r = \operatorname{rank}(A) \in [m-l; \min\{m; n+q-l\}]$ jeweils 10 Beispiele gerechnet. Insgesamt wurden so 12230 Beispiele ausgewertet. Die Elemente der Startränderungen sind analog zu [SS97b] im Intervall $[-5;+5]$ gleichverteilt. Bei Verwendung des Algorithmus 3.40, also im Fall $l > 0$, wurden anschließend die Spalten von ${}_{beg}\bar{D}^{(-1)}$ orthonormalisiert. Um den gewünschten Rang von A zu erhalten, wurde A mittels $\operatorname{rank}(A)$ Rang-1-Modifikationen der Nullmatrix bestimmt. Wie üblich erfolgte jede Rang-1-Modifikation mittels eines Tensorprodukts zweier Vektoren, wobei die Elemente jedes Vektors im Intervall $[-1;0]$ gleichverteilt sind.

Um den Algorithmus 3.40 ausführen zu können, wird in den Schritten S1.2 und S2.2 eine Rangbestimmung unterer Dreiecksmatrizen benötigt. Dabei wurde getestet, ob die Norm einer Spalte sehr klein bzw. ob sie kleiner als 10^{-8} und zusätzlich um 10^{-6} kleiner, also wesentlich kleiner, als die Norm der vorhergehenden Spalte ist. Wenn mindestens eine der beiden Bedingungen erfüllt waren, wurde diese und alle weiter rechts stehenden Spalten gleich dem Nullvektor gesetzt. In allen Beispielen wurde so der richtige Rang bestimmt. Der Wert von $\check{\sigma}$ wurde im Fall $0 < l < m$ gemäß Algorithmus 3.40 gewählt. Bei $l = m$

wurde bei $r > 0$ der Wert analog ausgewählt, was gerade $\|A\|$ entspricht, während bei $r = 0$, d. h., die Matrix A ist die Nullmatrix, $\check{\sigma} = 1$ gesetzt wurde. Im Fall $l = 0$, also bei Anwendung des Algorithmus aus Bemerkung 3.28, wurde für $\check{\sigma}$ die Norm der ersten Zeile von A verwendet. Alle Aussagen über $\check{\sigma}$ wurden bei einer Fehlertoleranz von 10^{-10} bestätigt.

Falls ${}_{neu}l = r - m + l \neq 0$ ist, wurde nach dem Startschritt Algorithmus 3.46 angewendet. Insbesondere ist bei $l = 0$ in jedem Fall $n + q - l = n + q > m$, vergleiche Definition 2.1, somit $r = m$, vergleiche Lemma 3.15 (ii) und Bemerkung 3.16 (iii), und damit ${}_{neu}l = 0$. Deshalb wird in diesem Fall nach dem Algorithmus aus Bemerkung 3.28 abgebrochen. Algorithmus 3.46 wird nach dem Schritt $i \to i+1$ beendet, falls die Bedingungen

$$
\begin{aligned}
\left\| \left(\left({}_{neu}\bar{D}^{(i)} \right)^T {}_{neu}\bar{D}^{(i+1)} \right)^{-1} \right\|^{-1} &= \cos\left(\measuredangle \left(\operatorname{im} {}_{neu}\bar{D}^{(i)}, \operatorname{im} {}_{neu}\bar{D}^{(i+1)} \right) \right) \geq 1 - tol\,, \\
\left\| \left(\left({}_{neu}T_0^{(i)} \right)^T {}_{neu}T_0^{(i+1)} \right)^{-1} \right\|^{-1} &= \cos\left(\measuredangle \left(\operatorname{im} {}_{neu}T_0^{(i)}, \operatorname{im} {}_{neu}T_0^{(i+1)} \right) \right) \geq 1 - tol
\end{aligned}
\tag{3.57}
$$

und

$$
\begin{aligned}
\sum_{j=1}^{{}_{neu}l} \sqrt{ \sum_{\iota=1}^{j-1} \left(\frac{\left(K^{(i+1)}\right)_{\iota,j}}{\left(K^{(i+1)}\right)_{j,j}} \right)^2 + \sum_{\iota=j+1}^{{}_{neu}l} \left(\frac{\left(K^{(i+1)}\right)_{\iota,j}}{\left(K^{(i+1)}\right)_{j,j}} \right)^2 } &\leq tol\,, \\
\sum_{j=1}^{{}_{neu}l} \sqrt{ \sum_{\iota=1}^{j-1} \left(\frac{\left(J^{(i+1)}\right)_{\iota,j}}{\left(J^{(i+1)}\right)_{j,j}} \right)^2 + \sum_{\iota=j+1}^{{}_{neu}l} \left(\frac{\left(J^{(i+1)}\right)_{\iota,j}}{\left(J^{(i+1)}\right)_{j,j}} \right)^2 } &\leq tol\,, \\
\frac{\sqrt{ \sum_{j=1}^{{}_{neu}l} \left(\left| \left(K^{(i+1)}\right)_{j,j} \right| - \left| \left(K^{(i)}\right)_{j,j} \right| \right)^2 }}{\sqrt{ \sum_{j=1}^{{}_{neu}l} \left(K^{(i+1)}\right)_{j,j}^2 }} &\leq tol\,, \\
\frac{\sqrt{ \sum_{j=1}^{{}_{neu}l} \left(\left| \left(J^{(i+1)}\right)_{j,j} \right| - \left| \left(J^{(i)}\right)_{j,j} \right| \right)^2 }}{\sqrt{ \sum_{j=1}^{{}_{neu}l} \left(J^{(i+1)}\right)_{j,j}^2 }} &\leq tol
\end{aligned}
\tag{3.58}
$$

für $tol = 10^{-10}$ allesamt erfüllt sind. (3.57) entspricht dabei in exakter Arithmetik

$$
\begin{aligned}
\left\| {}_{neu}\bar{D}^{(i+1)} - {}_{neu}\bar{D}^{(i)} \left({}_{neu}\bar{D}^{(i)} \right)^T {}_{neu}\bar{D}^{(i+1)} \right\|^2 &\leq 1 - (1 - tol)^2 = 2 * tol - tol^2 \quad \text{und} \\
\left\| {}_{neu}T_0^{(i+1)} - {}_{neu}T_0^{(i)} \left({}_{neu}T_0^{(i)} \right)^T {}_{neu}T_0^{(i+1)} \right\|^2 &\leq 1 - (1 - tol)^2 = 2 * tol - tol^2.
\end{aligned}
$$

Dieses Kriterium ist genau dann erfüllt, wenn sich die Bildräume von ${}_{neu}\bar{D}^{(i)}$ und ${}_{neu}T_0^{(i)}$ nur wenig ändern. Die Bedingungen in (3.58) bedeuten dagegen, daß die Nichtdiagonalelemente in $K^{(i+1)}$ bzw. $J^{(i+1)}$ klein gegenüber den entsprechenden Diagonalelementen sind

und daß sich diese Diagonalelemente nur wenig ändern. In den meisten betrachteten Fällen ergab sich das gleiche Abbruchkriterium, wenn in den letzten beiden Bedingungen in (3.58) die Betragsstriche weggelassen wurden. Bis auf einen Fall bei $n+q-l=4$ und $m=l=r=2$ traten Unterschiede lediglich bei $m=l$ und $r=1$ auf. In diesen Fällen ist ${}_{neu}l=1$. Diese Unterschiede bedeuten gerade, daß $K^{(i)}$ und $K^{(i+1)}$ bzw. $J^{(i)}$ und $J^{(i+1)}$ sich betragsmäßig kaum unterschieden, aber entgegengesetztes Vorzeichen hatten. Bei $n+q-l=4$ und $m=l=r=2$ hatten zumindest die zum größeren Singulärwert gehörenden Diagonaleinträge von $J^{(i)}$ und $J^{(i+1)}$ entgegengesetzte Vorzeichen, während dies bei $K^{(i)}$ und $K^{(i+1)}$ nicht der Fall war. Zur näheren Betrachtung wurden die relativen Häufigkeiten dieser Vorzeichenwechsel bei $m=l$, $r=1$ und $n\neq 30$ untersucht. Bei 240 betrachteten Fällen traten beim Abbruchschritt 119 Vorzeichenwechsel bei K, 138 Vorzeichenwechsel bei J und davon 58 gleichzeitige Vorzeichenwechsel auf. Bis auf den Wert bei J weist das auf Wahrscheinlichkeit 1/2 und Unabhängigkeit hin. Die Fälle $m=l$ und $r=1$ sind gerade die Fälle für $r-m+l\neq 0$, in denen bereits im Algorithmus 3.40 unabhängig von den Startwerten die gesuchte Lösung berechnet wird. Deshalb wird stets nur ein Schritt des Algorithmus 3.46 in diesen Fällen durchgeführt. Bei dem erwähnten Fall mit $n+q-l=4$ und $m=l=r=2$ wurden 2 Schritte des Algorithmus 3.46 ausgeführt.

Wegen Algorithmus 3.40 gilt in exakter Arithmetik $\ker A = \operatorname{im}{}_sT_0^{(0)} \perp \operatorname{im}{}_rT_0^{(0)}$ und $\operatorname{im}{}_{neu}\bar{D}^{(0)} = \operatorname{im}{}_r\bar{D}^{(0)} \perp \operatorname{im}{}_s\bar{D}^{(0)} = \ker\left[A^T \quad {}_sT_0^{(0)}\right]$. Somit gilt in exakter Arithmetik $\operatorname{im}{}_{neu}\bar{D}^{(0)} = \operatorname{span}\{M_{neu\,m,m}\,u^i, i=1,2,\ldots,r\}$. Analoge Aussagen gelten mit einer etwas anderen Begründung auch für ${}_{neu}T_0^{(0)}$. Für $r=1$ folgt wegen $l\leq m$, vergleiche (2.3), und $r\geq m-l$ von $r-m+l=0$ und $m=l$ mindestens eine Gleichung. Daraus ergibt sich, daß in exakter Arithmetik bereits Algorithmus 3.40 die gesuchten Singulärvektoren und den gesuchten Singulärwert liefert.

Entsprechend (3.49) gemäß der Folgerung 3.55 sind die Konditionszahlen der Matrizen B_i aus den Schritten S1.1 und S2.1 des Algorithmus 3.46 monoton fallend. Eine entsprechende Aussage für den Algorithmus 3.40 gilt jedoch nicht. D. h., $\operatorname{cond} B({}_{beg}\bar{D}^{(-1)}, {}_{beg}T_0^{(-1)}, ({}_{beg}K^{(-1)})^T)$ kann zwar größer als die Konditionszahl der entsprechenden Matrix aus S2.1 und als $\operatorname{cond} B({}_{neu}\bar{D}^{(0)}, {}_{neu}T_0^{(0)}, ({}_{neu}K^{(0)})^T)$ sein, dies muß jedoch nicht der Fall sein. Bei den betrachteten Beispielen war der größte relative Zuwachs bei einem Fall mit $A\in\mathbb{R}^{8\times 4}$, $l=7$ und $r=1$ mit $\operatorname{cond} B({}_{beg}\bar{D}^{(-1)}, {}_{beg}T_0^{(-1)}, ({}_{beg}K^{(-1)})^T)\approx 53$, während die Konditionszahl der entsprechenden Matrix aus S2.1 den Wert ≈ 5000 besaß. Der größte absolute Zuwachs geschah dagegen in einem Beispiel mit $A\in\mathbb{R}^{25\times 25}$, $l=5$ und $r=20$, wobei $\operatorname{cond} B({}_{beg}\bar{D}^{(-1)}, {}_{beg}T_0^{(-1)}, ({}_{beg}K^{(-1)})^T)\approx 19000$, während die Konditionszahl der entsprechenden Matrix aus S2.1 den Wert ≈ 937000 hatte.

Nach Beendigung des Algorithmus 3.46 wurde mit den berechneten genäherten Singulärwerten und -vektoren und $\check{\sigma}$ gemäß Bemerkung 3.11 (iii) eine genäherte optimale Ränderung berechnet. Deren Konditionszahl wurde mit der theoretischen optimalen Konditionszahl verglichen. Der absolute Fehler lag dabei im Intervall $[-1.22*10^{-5}; 7.65*10^{-5}]$ mit dem Mittelwert bei $\approx 1.79*10^{-8} = 1.79e-8$ und der relative Fehler im Intervall $[-7*10^{-12}; 1.67*10^{-5}]$. Die negativen Werte sind nach der Interlacing Property, vergleiche (3.2), nicht möglich. Sie beruhen sicherlich auf Fehler bei der Ermittlung derjenigen Singulärwerte, die bei der Berechnung der optimalen und der genähert optimalen Konditionszahl verwendet wurden. Die bezüglich des relativen Fehlers schlechteste Konditionszahl der genäherten optimalen Ränderung wurde dabei in dem Beispiel erreicht, bei dem bezüglich der Ausgangskonditions-

zahl die größte relative Verbesserung erzielt wurde. Die optimale Konditionszahl beträgt in diesem Beispiel 1, während $\operatorname{cond} B({}_{beg}\bar{D}^{(-1)}, {}_{beg}T_0^{(-1)}, ({}_{beg}K^{(-1)})^T) \approx 10^7$. Somit wurden ziemlich gute Näherungen der optimal geänderten Matrizen gewonnen.

Zum Schluß dieses Unterabschnitts folgt noch eine Übersicht über die im Algorithmus 3.46 benötigten Schritte für verschiedene Werte von ${}_{neu}m$ und ${}_{neu}l$. Dabei werden die 3090 Beispiele mit ${}_{neu}l = 0$ nicht berücksichtigt, da dort Algorithmus 3.46 nicht verwendet wird. Dabei sei i_{mittel} die mittlere, i_{max} die größte und i_{min} die kleinste benötigte Schrittzahl bei den jeweiligen Beispielen.

Beispiele	Beispielanzahl	i_{mittel}	i_{max}	i_{min}
${}_{neu}m \in [2;6]$	360	12.04	126	1
${}_{neu}m \in [7;11]$	1660	26.94	720	1
${}_{neu}m \in [12;16]$	1230	31.64	530	1
${}_{neu}m \in [17;21]$	1340	39.05	429	1
${}_{neu}m \in [22;26]$	1620	46.42	1739	1
${}_{neu}m \in [27;31]$	1900	48.08	1063	1
${}_{neu}m \in [32;36]$	1030	43.62	1120	1
${}_{neu}l = 1$	2730	10.55	313	1
${}_{neu}l = 2$	2150	25.66	430	2
${}_{neu}l = 3$	1590	43.54	1624	5
${}_{neu}l = 4$	1180	61.26	1739	7
${}_{neu}l = 5$	790	73.59	584	16
${}_{neu}l = 6$	490	89.91	1259	15
${}_{neu}l = 7$	210	114.82	1063	26

Die mittleren Schrittzahlen liegen durchaus im akzeptablen Bereich. Somit ist die vorgeschlagene Algorithmenkombination sinnvoll anwendbar. Die mittleren Schrittzahlen wachsen kaum mit wachsendem ${}_{neu}m$, aber deutlicher mit wachsendem ${}_{neu}l$.

3.3 Ausblick auf das Newtonverfahren

Wie im Abschnitt 3.2 beschrieben wird, ist die verallgemeinerte inverse Iteration für fast alle Startwerte ein konvergentes Verfahren. Allerdings ist die Konvergenz nur linear. Wenn für einen Punkt (x, λ, α) eine gute Näherung für die Singulärvektoren zu den kleinsten Singulärwerten bekannt ist, ist dies meist auch eine brauchbare Näherung für die Singulärvektoren zu den kleinsten Singulärwerten für Punkte aus einer Umgebung von (x, λ, α). Eine Voraussetzung ist dabei die stetige Differenzierbarkeit von F. Dann hängen die Singulärwerte und die zugehörigen Singulärvektoren stetig vom Argument der Jacobimatrix ab, vergleiche [GVL89, P 7.1.6, S. 340] und Bemerkung 3.3 (iv). Jedoch können durch Veränderungen in der Reihenfolge der Singulärwerte bei Änderung der Argumente (x, λ, α) andere Singulärwerte zu den kleinsten werden. Damit gehören auch andere Singulärvektoren zu den kleinsten Singulärwerten. Deshalb hängen die Singulärvektoren zu den jeweils kleinsten Singulärwerten nicht überall stetig von (x, λ, α) ab. Somit ist weiterhin vorauszusetzen, daß gewisse Singulärwerte ihre Reihenfolge nicht vertauschen. Wenn dies erfüllt ist, kann für die Berechnung der Singulärvektoren zu den kleinsten Singulärwerten ein lokal

konvergentes Verfahren wie z. B. das Newtonverfahren verwendet werden. Als Startränderung kann die berechnete Ränderung für einen Punkt aus einer Umgebung des aktuellen Punktes (x, λ, α) verwendet werden. Wenn (x, λ, α) eine Iterierte des Newtonverfahrens bei der Berechnung eines singulären Punktes ist, kann beispielsweise in vielen Fällen die Ränderung für die vorhergehende Iterierte als Startränderung zur Berechnung der optimalen Ränderung im aktuellen Punkt verwendet werden. Wenn das Newtonverfahren zur Berechnung der zu den kleinsten Singulärwerten gehörenden Singulärvektoren angewendet werden kann, ist die Konvergenz unter den üblichen Voraussetzungen bekanntlich quadratisch.

Deshalb sollen in diesem Abschnitt einige Bemerkungen zur Verwendung des Newtonverfahrens zur Berechnung solcher Singulärvektoren erfolgen. Newtonverfahren zur Berechnung von Eigenvektoren sind bekannt, vergleiche z. B. [SL97] und [LST98]. Die dort beschriebenen Verfahren lassen sich beispielsweise auf $A^T A$, $A\,A^T$, $\begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix}$ bzw. $\begin{bmatrix} 0 & A^T \\ A & 0 \end{bmatrix}$ anwenden. Dabei ist jedoch zu beachten, daß bei regulären Matrizen

$$\operatorname{cond} A^T A = \operatorname{cond} A\,A^T = (\operatorname{cond} A)^2 \quad \text{und} \quad \operatorname{cond} A = \operatorname{cond} \begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix} = \operatorname{cond} \begin{bmatrix} 0 & A^T \\ A & 0 \end{bmatrix}$$

gilt. Außerdem liefert die direkte Verwendung von $A^T A$ bzw. $A\,A^T$ nur die Rechts- bzw. Linkssingulärvektoren von A. Deshalb sind die anderen Varianten zu bevorzugen.

Bei der Übertragung der dort verwendeten Eigenwertgleichung wird ausgenutzt, daß für die Lösungen gerade $A V_4 - U_4 \Sigma_4 = 0$ und $A^T U_4 - V_4 \Sigma_4^T = 0$ gilt. Bei der Konstruktion des zu linearisierenden Systems ist zu beachten, daß bei jedem zu berechnenden Eigenwert von $\begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix}$ der gesamte Eigenraum zu berechnen ist. Sonst ist die Jacobimatrix für die Lösung nicht regulär. Wenn insbesondere ein Eigenvektor für den eventuell vorhandenen Eigenwert Null, d. h. ein Nullraumvektor, zu bestimmen ist, sind sämtliche Nullraumvektoren zu berechnen. Eigenwerte Null treten genau dann auf, wenn A nicht regulär z. B. nicht quadratisch ist. Für alle Eigenwerte ungleich Null brauchen dagegen die zu dem mit (-1) multiplizierten Eigenwert gehörenden Eigenvektoren nicht mit berechnet werden. Um bei diesen Eigenwerten keine Probleme zu bekommen, wird wieder $\kappa := \sigma_{m-l+1}/\sigma_{m-l} < 1$ vorausgesetzt.

Im Folgenden sollen nun einige erweiterte Systeme für das Newtonverfahren erwähnt werden. Dabei werden die Bezeichnungen aus Abschnitt 3.2 verwendet. Die Dimensionen stimmen ja überein. Es sind lediglich kleine Unterschiede in der Berechnungsvorschrift zu beachten. Der Nachweis der Voraussetzungen für die quadratische Konvergenz des Newtonverfahrens erfolgt nicht. Entsprechend [LST98, Abschnitt 2] kann das Blocknewtonverfahren mit einem Rayleigh-Ritz-Verfahren kombiniert werden.

Falls A regulär ist:

$$\begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} \bar{D} \\ T_0 \end{bmatrix} - \begin{bmatrix} \bar{D} \\ T_0 \end{bmatrix} J = 0\,, \quad \text{d. h.,} \quad A\,T_0 - \bar{D}\,J = 0 \quad \text{und} \quad A^T \bar{D} - T_0\,J = 0$$

mit der zusätzlichen Bedingung $(\bar{D}^{(i)})^T \bar{D} + (T_0^{(i)})^T T_0 - I_l = 0$ entsprechend [LST98, (P_k)] bzw. $(\bar{D}^{(i)})^T \bar{D} + (T_0^{(i)})^T T_0 - 2I_l = 0$ analog zu den Algorithmen 3.27, 3.44 und 3.46.

Im allgemeinen Fall von A:

$$A\,T_0 - \bar{D}\,J = 0 \quad \text{und} \quad A^T \bar{D} - T_0\,K = 0 \tag{3.59}$$

mit den zusätzlichen Bedingungen $(\bar{D}^{(i)})^T\bar{D} - I_l = 0$ und $(T_0^{(i)})^T T_0 - I_{n-m+q} = 0$. (3.59) kann dabei auch als die nichttrivialen Teilblöcke von

$$\begin{bmatrix} A\,T_0 - \bar{D}\,J & 0 \\ 0 & A^T\bar{D} - T_0\,K \end{bmatrix} = \begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} 0 & \bar{D} \\ T_0 & 0 \end{bmatrix} - \begin{bmatrix} 0 & \bar{D} \\ T_0 & 0 \end{bmatrix} \begin{bmatrix} 0 & K \\ J & 0 \end{bmatrix} = 0 \qquad \text{bzw.}$$

$$\begin{bmatrix} 0 & A\,T_0 - \bar{D}\,J \\ A^T\bar{D} - T_0\,K & 0 \end{bmatrix} = \begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} \bar{D} & 0 \\ 0 & T_0 \end{bmatrix} - \begin{bmatrix} 0 & \bar{D} \\ T_0 & 0 \end{bmatrix} \begin{bmatrix} K & 0 \\ 0 & J \end{bmatrix} = 0$$

definiert werden.

Eine direkte Übertragung von [LST98, (P_k)] ergibt im Fall einer quadratischen Matrix A dagegen:

$$\begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} \bar{D} & \bar{D} \\ T_0 & -T_0 \end{bmatrix} - \begin{bmatrix} \bar{D} & \bar{D} \\ T_0 & -T_0 \end{bmatrix} \begin{bmatrix} J & 0 \\ 0 & -J \end{bmatrix} = \begin{bmatrix} A\,T_0 - \bar{D}\,J & -(A\,T_0 - \bar{D}\,J) \\ A^T\bar{D} - T_0\,J & A^T\bar{D} - T_0\,J \end{bmatrix} = 0$$

mit weiteren Bedingungen. Dabei brauchen jedoch die beiden doppelt auftretenden Bedingungen nur einfach verwendet werden.

Bei der Übertragung des Systems [LST98, (P_k)] sind auch noch weitere Varianten denkbar.

Falls die gesuchten Eigenwerte von $\begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix}$ einfach sind, können auch die einzelnen Eigenvektoren mit den Verzweigungspunktalgorithmen 3 (MGRE) bzw. 9 (EMGRE) aus [SL97] berechnet werden. Für die Einfachheit der gesuchten Eigenwerte ist die Einfachheit der entsprechenden Singulärwerte von A notwendig. Die gesuchten Singulärwerte ergeben sich aus den Beträgen der Eigenwerte, während die Singulärvektoren bis auf das Vorzeichen die entsprechenden Teile der Eigenvektoren sind.

Kapitel 4

Ein zweistufiges Newtonverfahren

In diesem Kapitel sollen die erweiterten Systeme gemäß Definition 2.3 in der speziellen Gestalt (2.10), die für $l = 0$ der Gestalt (2.8) entspricht, gelöst werden. Dafür wird das überlinear konvergente Newtonverfahren verwendet. Bei Verwendung der charakterisierenden Funktionen aus Abschnitt 2.4, speziell aus den Unterabschnitten 2.4.2 und 2.4.3, werden in jedem Fall Ableitungen der Funktion F mindestens zweiter Ordnung benötigt. In einigen früheren Arbeiten, z. B. [PS81] und [Pön87] wurden diese Ableitungen höherer Ordnung durch Differenzenquotienten angenähert. Hier soll nun dargestellt werden, wie sich die erweiterten Systeme mittels des Newtonverfahrens lösen lassen, wobei alle benötigten Ableitungen durch Anwendung der Automatischen Differentiation bestimmt werden. Dabei wird die Vorgehensweise für gewisse Spezialfälle von $m = n$, $p = q = 1$ aus [PSS99] auf geeignete erweiterte Systeme für beliebige singuläre Punkte verallgemeinert. Wie in jener Arbeit wird auch hier die Struktur von $G(y)$ in einem zweistufigen Newtonverfahren ausgenutzt. Diese Vorgehensweise geht auf Pönisch/Schwetlick [PS81] zurück. Sie wurde ebenfalls u. a. in [Pön87], [Pön90], [Pön94], [Sch94], [PSS97b] und [AS97] verwendet. In [PSS99] wird außerdem der Zusammenhang zwischen den verwendeten Richtungsableitungen von f und gewissen Ableitungen von $\boldsymbol{\vartheta} = \boldsymbol{\lambda}$ dargestellt. Dies wird in der vorliegenden Arbeit verallgemeinert. Die Jacobimatrix von F, f und die benötigten Richtungsableitungen von f werden mit Hilfe der Automatischen Differentiation bestimmt. Der Rechenaufwand des zweistufigen Newtonverfahrens ist dabei wegen der speziellen Gestalt von f kleiner als beim einfachen Newtonverfahren. Dieses Kapitel ist eine leicht überarbeitete Version von [SW99].

Im folgenden Abschnitt 4.1 wird der sich aus dem zweistufigen Newtonverfahren ergebende Algorithmus hergeleitet. Anschließend wird die Konvergenz gezeigt. Einige Bemerkungen zur Anwendung der Automatischen Differentiation in diesem Algorithmus folgen im Abschnitt 4.2. Dieses Verfahren wurde für Spezialfälle von $m = n$, $p = q = 1$ implementiert. Einige Ergebnisse der mit diesem Programm erfolgten Berechnungen sind im Abschnitt 4.3 enthalten.

4.1 Algorithmische Beschreibung

Es sei $y^{(i)} = y = (x, \lambda, \alpha, \mu) = (x, \vartheta, \tau)$ die aktuelle Iterierte. Dann ergibt sich die nächste Iterierte $y^{(i+1)} = y^+ = (x^+, \lambda^+, \alpha^+, \mu^+) = (x^+, \vartheta^+, \tau^+)$ aus der Linearisation von (2.60),

d. h.,

$$\partial \boldsymbol{F}(x,\vartheta,\tau)\begin{pmatrix} x^+ - x \\ \vartheta^+ - \vartheta \\ \tau^+ - \tau \end{pmatrix} = -\boldsymbol{F}(x,\vartheta,\tau), \tag{4.1}$$

$$\partial \boldsymbol{f}(x,\vartheta,\tau)\begin{pmatrix} x^+ - x \\ \vartheta^+ - \vartheta \\ \tau^+ - \tau \end{pmatrix} = -\boldsymbol{f}(x,\vartheta,\tau). \tag{4.2}$$

Wegen (2.48) und der Glattheit von F ist $\partial \boldsymbol{F}(x,\vartheta,\tau)$ in einer Umgebung des singulären Punktes $y^* = (x^*,\lambda^*,\alpha^*,0) = (x^*,\vartheta^*,\tau^*)$ zeilenregulär. Somit ist

$$\dim\ker \partial \boldsymbol{F}(x,\vartheta,\tau) = n+p+k+l-m$$

und es gilt wegen der ersten Ableitung von der ersten Zeile von (2.55) nach $(\boldsymbol{\xi},\tau)$

$$\ker \partial \boldsymbol{F}(x,\vartheta,\tau) = \operatorname{im}\begin{bmatrix} \partial_{\boldsymbol{\xi}}\boldsymbol{x}(\boldsymbol{\xi},\tau,c) & \partial_\tau \boldsymbol{x}(\boldsymbol{\xi},\tau,c) \\ \partial_{\boldsymbol{\xi}}\boldsymbol{\vartheta}(\boldsymbol{\xi},\tau,c) & \partial_\tau \boldsymbol{\vartheta}(\boldsymbol{\xi},\tau,c) \\ 0 & I_{p+k+l-q} \end{bmatrix} \tag{4.3}$$

mit $(c,\boldsymbol{\xi}) = (\boldsymbol{F}(x,\vartheta,\tau), \boldsymbol{T}(x-x^*,\vartheta-\vartheta^*))$, vergleiche (2.61), und

$$\begin{bmatrix} V & W \\ 0 & I_{p+k+l-q} \end{bmatrix} := \begin{bmatrix} \partial_{\boldsymbol{\xi}}\boldsymbol{x}(\boldsymbol{\xi},\tau,c) & \partial_\tau \boldsymbol{x}(\boldsymbol{\xi},\tau,c) \\ \partial_{\boldsymbol{\xi}}\boldsymbol{\vartheta}(\boldsymbol{\xi},\tau,c) & \partial_\tau \boldsymbol{\vartheta}(\boldsymbol{\xi},\tau,c) \\ 0 & I_{p+k+l-q} \end{bmatrix}. \tag{4.4}$$

Damit ist die allgemeine Lösung des Systems (4.1) wegen der ersten Ableitung der ersten Zeile von (2.55) nach c

$$y^+ - y = \begin{pmatrix} x^+ - x \\ \vartheta^+ - \vartheta \\ \tau^+ - \tau \end{pmatrix} = \begin{bmatrix} V & W \\ 0 & I_{p+k+l-q} \end{bmatrix}\gamma - \begin{bmatrix} S \\ 0 \end{bmatrix} \quad \text{mit} \tag{4.5}$$

$$\begin{bmatrix} S \\ 0 \end{bmatrix} := \begin{bmatrix} \partial_c \boldsymbol{x}(\boldsymbol{\xi},\tau,c) \\ \partial_c \boldsymbol{\vartheta}(\boldsymbol{\xi},\tau,c) \\ 0 \end{bmatrix} \boldsymbol{F}(x,\vartheta,\tau) \tag{4.6}$$

und $(c,\boldsymbol{\xi}) = (\boldsymbol{F}(x,\vartheta,\tau), \boldsymbol{T}(x-x^*,\vartheta-\vartheta^*))$, vergleiche (2.61). Zur Berechnung der impliziten Ableitungen wird dabei die Matrix

$$\boldsymbol{B}(x,\vartheta,\tau) := \begin{bmatrix} \partial_{(x,\vartheta)}\boldsymbol{F}(x,\vartheta,\tau) \\ \boldsymbol{T} \end{bmatrix} \tag{4.7}$$

verwendet, die wegen der Regularität von $\boldsymbol{T}\begin{bmatrix} V_s \\ 0 \end{bmatrix}$, vergleiche Bemerkung 2.36 (ii), in der betrachteten Umgebung des singulären Punktes (x^*,ϑ^*,τ^*) regulär ist. Wenn diese allgemeine Lösung von (4.1) in (4.2) eingesetzt wird, folgt

$$\partial \boldsymbol{f}(x,\vartheta,\tau)\begin{bmatrix} V & W \\ 0 & I_{p+k+l-q} \end{bmatrix}\gamma = \partial \boldsymbol{f}(x,\vartheta,\tau)\begin{bmatrix} S \\ 0 \end{bmatrix} - \boldsymbol{f}(x,\vartheta,\tau) \in \mathbb{R}^{n+p+k+l-m} \tag{4.8}$$

mit $(c,\boldsymbol{\xi}) = (\boldsymbol{F}(x,\vartheta,\tau), \boldsymbol{T}(x-x^*,\vartheta-\vartheta^*))$, vergleiche (2.61). Daraus kann γ bestimmt werden. Deshalb ergibt sich folgender Algorithmusschritt für die Berechnung von $y^{(i+1)} := y^+$ aus $y := y^{(i)}$.

Algorithmus 4.1.

1. Berechne $\boldsymbol{F}(y), \partial \boldsymbol{F}(y)$ mittels Automatischem Differenzieren und bilde $\boldsymbol{B}(y)$ entsprechend (4.7).
2. Berechne $\begin{bmatrix} V & W & S \end{bmatrix} \in \mathbb{R}^{(n+q)\times(n+p+k+l-m+1)}$ an der aktuellen Stelle y mittels
$$\boldsymbol{B}(y) \begin{bmatrix} V & W & S \end{bmatrix} = \begin{bmatrix} 0 & -\partial_\tau \boldsymbol{F}(y) & \boldsymbol{F}(y) \\ I_{n+q-m} & 0 & 0 \end{bmatrix}.$$
3. Berechne $\boldsymbol{f}(y), \partial \boldsymbol{f}(y) \begin{bmatrix} V \\ 0 \end{bmatrix}, \partial \boldsymbol{f}(y) \begin{bmatrix} W \\ I_{p+k+l-q} \end{bmatrix}$ und $\partial \boldsymbol{f}(y) \begin{bmatrix} S \\ 0 \end{bmatrix}$ mittels Automatischem Differenzieren.
4. Löse (4.8) für $\gamma \in \mathbb{R}^{n+p+k+l-m}$.
5. Berechne $\begin{pmatrix} x^+ \\ \vartheta^+ \end{pmatrix} := \begin{pmatrix} x \\ \vartheta \end{pmatrix} + \begin{bmatrix} V & W \end{bmatrix} \gamma - S, \quad \tau^+ = \tau + \begin{bmatrix} 0 & I_{p+k+l-q} \end{bmatrix} \gamma.$

Für diesen Algorithmus gilt folgender Konvergenzsatz:

Satz 4.2.
Die Funktion F und damit auch die Funktion $\boldsymbol{F}$ sei in einer Umgebung des singulären Punktes $(x^*, \lambda^*, \alpha^*)$ bzw. $y^* = (x^*, \vartheta^*, \tau^*)$ hinreichend glatt, so daß G aus (2.60) in dieser Umgebung Lipschitz-stetig differenzierbar ist. Dann ist Algorithmus 4.1 unbeschränkt durchführbar für alle Startwerte aus einer hinreichend kleinen Kugelumgebung von y^* und die durch Algorithmus 4.1 erzeugte Folge $\{y^{(i)}\}$ konvergiert mindestens Q-quadratisch gegen y^*.

Der Beweis dieses Satzes folgt aus [Sch79], Satz 5.1.2, unter Beachtung der Regularität von $\partial G(y^*)$, siehe Definition 2.3.

Bemerkung 4.3.

(i) Da sich f und damit auch $\boldsymbol{f}$ gemäß Abschnitt 2.4 allein mittels Ableitungen von $\boldsymbol{\vartheta}$ darstellen läßt, hängt nach dem Satz über die impliziten Funktionen die Glattheit von $\boldsymbol{f}$ und damit auch von G nur von der Glattheit von $\boldsymbol{F}$ und somit nur von der Glattheit von F ab. Falls insbesondere die höchste in f auftretende Ableitung von $\boldsymbol{\vartheta}$ die i-te Ableitung ist, so ist G lokal Lipschitz-stetig differenzierbar, falls F $(i+1)$-mal lokal Lipschitz-stetig differenzierbar ist.

(ii) Da $\mu^* = 0$ bekannt ist, wird in der Praxis häufig in jedem Schritt i nicht $\mu^{(i+1)} := \mu^+$, sondern $\mu^{(i+1)} := 0$ gesetzt, siehe z. B. [AS97] und [PSS97b]. In den üblichen Normen, z. B. in der euklidischen Norm, konvergiert auch diese Folge lokal Q-quadratisch, da für diese Normen gilt:
$$\begin{aligned} \|(x^{(i+1)}, \lambda^{(i+1)}, \alpha^{(i+1)}) - (x^*, \lambda^*, \alpha^*)\| &\le \|(x^{(i+1)}, \lambda^{(i+1)}, \alpha^{(i+1)}, \mu^+) - (x^*, \lambda^*, \alpha^*, 0)\| \\ &\le Q\|(x^{(i)}, \lambda^{(i)}, \alpha^{(i)}, 0) - (x^*, \lambda^*, \alpha^*, 0)\|^2 \end{aligned}$$

(iii) Bei praktischen Anwendungen werden gelegentlich nach jedem Schritt i die Spalten von V orthonormiert und die so erhaltene Matrix $\bar{T}^{(i)}$ im nächsten Iterationsschritt $i+1$ in der Matrix $\boldsymbol{B}(y^{(i+1)})$ als Matrix $\boldsymbol{T}^T$ verwendet. Offensichtlich ist $V, \bar{T}^{(i)} \in \mathbb{R}^{(n+q)\times(n+q-m)}$ und $\boldsymbol{T} \in \mathbb{R}^{(n+q-m)\times(n+q)}$, so daß dies möglich ist. $\bar{T}^{(i)}$ entspricht dabei $(\boldsymbol{T}^{(1)})^T$ aus dem Algorithmus in der Bemerkung 3.28.

4.2 Anwendung des Automatischen Differenzierens

Die Grundlagen des Automatischen Differenzierens wurden beispielsweise von Griewank [Gri89] beschrieben. Die Berechnung von $\boldsymbol{F}(y)$ und $\partial \boldsymbol{F}(y)$ mittels Automatischer Differentiation ist Standard, vergleiche die function- und die jacobian-Funktion in [GJU96]. Deshalb soll hier nicht weiter darauf eingegangen werden. Im folgenden soll deshalb Schritt 3 im Algorithmus 4.1 näher untersucht werden.

Da f und somit auch $\boldsymbol{f}$ sich allein mittels Ableitungen von $\boldsymbol{\vartheta}$ darstellen läßt und wegen (2.61) $(\boldsymbol{x}(\boldsymbol{\xi},\tau,c),\boldsymbol{\vartheta}(\boldsymbol{\xi},\tau,c)) = (x,\vartheta)$ gesetzt werden kann, lassen sich auch

$$
\begin{aligned}
\partial \boldsymbol{f}(y) \begin{bmatrix} V \\ 0 \end{bmatrix} &= \partial_{(x,\vartheta)} \boldsymbol{f}(\boldsymbol{x}(\boldsymbol{\xi},\tau,c),\boldsymbol{\vartheta}(\boldsymbol{\xi},\tau,c),\tau) \begin{bmatrix} \partial_{\boldsymbol{\xi}} \boldsymbol{x}(\boldsymbol{\xi},\tau,c) \\ \partial_{\boldsymbol{\xi}} \boldsymbol{\vartheta}(\boldsymbol{\xi},\tau,c) \end{bmatrix} \\
&= d_{\boldsymbol{\xi}} \boldsymbol{f}(\boldsymbol{x}(\boldsymbol{\xi},\tau,c),\boldsymbol{\vartheta}(\boldsymbol{\xi},\tau,c),\tau), \\
\partial \boldsymbol{f}(y) \begin{bmatrix} W \\ I_{p+k+l-q} \end{bmatrix} &= \partial \boldsymbol{f}(\boldsymbol{x}(\boldsymbol{\xi},\tau,c),\boldsymbol{\vartheta}(\boldsymbol{\xi},\tau,c),\tau) \begin{bmatrix} \partial_{\tau} \boldsymbol{x}(\boldsymbol{\xi},\tau,c) \\ \partial_{\tau} \boldsymbol{\vartheta}(\boldsymbol{\xi},\tau,c) \\ I_{p+k+l-q} \end{bmatrix} \\
&= d_{\tau} \boldsymbol{f}(\boldsymbol{x}(\boldsymbol{\xi},\tau,c),\boldsymbol{\vartheta}(\boldsymbol{\xi},\tau,c),\tau) \qquad \text{und} \\
\partial \boldsymbol{f}(y) \begin{bmatrix} S \\ 0 \end{bmatrix} &= \partial_{(x,\vartheta)} \boldsymbol{f}(\boldsymbol{x}(\boldsymbol{\xi},\tau,c),\boldsymbol{\vartheta}(\boldsymbol{\xi},\tau,c),\tau) \begin{bmatrix} \partial_{c} \boldsymbol{x}(\boldsymbol{\xi},\tau,c) \\ \partial_{c} \boldsymbol{\vartheta}(\boldsymbol{\xi},\tau,c) \end{bmatrix} \boldsymbol{F}(x,\vartheta,\tau) \\
&= d_{c} \boldsymbol{f}(\boldsymbol{x}(\boldsymbol{\xi},\tau,c),\boldsymbol{\vartheta}(\boldsymbol{\xi},\tau,c),\tau)\, \boldsymbol{F}(x,\vartheta,\tau)
\end{aligned} \tag{4.9}
$$

allein mittels Richtungsableitungen der implizit definierten Funktion $\boldsymbol{\vartheta}$ darstellen. Dies bedeutet, daß zur Berechnung von $\boldsymbol{f}$ und der benötigten Richtungsableitungen von $\boldsymbol{f}$ nur Richtungsableitungen von $\boldsymbol{\vartheta}$ gebraucht werden. Für den Spezialfall $n = m$, $p = q = 1$, $\lambda = \vartheta$ wurden die entsprechende Berechnungen in [PSS99] beschrieben. Weiterhin sind Berechnungsmöglichkeiten für Richtungsableitungen expliziter Funktionen mittels Automatischer Differentiation bekannt. In [GUW97] wird dargestellt, warum zur Berechnung dieser Richtungsableitungen univariate Taylorentwicklungen verwendet werden sollten und wie sich aus den erhaltenen Taylorkoeffizienten die gesuchten Richtungsableitungen bestimmen lassen. Die dort beschriebene Vorgehensweise soll nun auf den hier behandelten Fall unter Verallgemeinerung der Ergebnisse von [PSS99] übertragen werden.

Analog zu [PSS99] sei $s := (x,\vartheta)$, $\varrho := (\boldsymbol{\xi},\tau,c)$ und

$$
\Phi(s,\varrho) := \begin{bmatrix} \boldsymbol{F}(x,\vartheta,\tau) - c \\ \boldsymbol{T}(x - x^*,\vartheta - \vartheta^*) - \boldsymbol{\xi} \end{bmatrix}. \tag{4.10}
$$

Aus der ersten Zeile von (2.55) folgt dann

$$
\Phi(\boldsymbol{s}(\varrho),\varrho) = \begin{bmatrix} \boldsymbol{F}(\boldsymbol{x}(\boldsymbol{\xi},\tau,c),\boldsymbol{\vartheta}(\boldsymbol{\xi},\tau,c),\tau) - c \\ \boldsymbol{T}(\boldsymbol{x}(\boldsymbol{\xi},\tau,c) - x^*,\boldsymbol{\vartheta}(\boldsymbol{\xi},\tau,c) - \vartheta^*) - \boldsymbol{\xi} \end{bmatrix} \equiv 0. \tag{4.11}
$$

Wegen $\vartheta = N^T_{n+q,q} s$ können deshalb alle benötigten Ableitungen von $\boldsymbol{\vartheta}$ aus den entsprechenden Ableitungen von $\boldsymbol{s}$ gewonnen werden. Somit sind Taylorkoeffizienten der implizit definierten Funktion $\boldsymbol{s}(\varrho + a)$ für geeignete Richtungen a zu bestimmen. Für hinreichend glatte Funktionen F in einer Umgebung von (x^*,λ^*,α^*) und damit hinreichend glatte Funktionen Φ in einer Umgebung von $(x^*,\vartheta^*,0,\tau^*,0)$ gilt somit für hinreichend kleine $a \in \mathbb{R}^{n+p+k+l}$

wegen (4.11)

$$0 \equiv \Phi(\boldsymbol{s}(\varrho+a),\varrho+a) \equiv \Phi\left(\sum_{i=0}^{j}\frac{1}{i!}\partial^i\boldsymbol{s}(\varrho)[a]^i + O(\|a\|^{j+1}),\varrho+a\right).$$

Wenn nun Φ^{j+1} gemäß

$$\Phi\left(\sum_{i=0}^{j}\frac{1}{i!}\partial^i\boldsymbol{s}(\varrho)[a]^i,\varrho+a\right) =: \Phi^{j+1}[a]^{j+1} + O(\|a\|^{j+2}), \qquad j=0,1,\ldots \tag{4.12}$$

definiert wird, ergibt sich

$$\begin{aligned}
\Phi\left(\sum_{i=0}^{j+1}\frac{1}{i!}\partial^i\boldsymbol{s}(\varrho)[a]^i,\varrho+a\right) &= \Phi\left(\sum_{i=0}^{j}\frac{1}{i!}\partial^i\boldsymbol{s}(\varrho)[a]^i + \frac{1}{(j+1)!}\partial^{j+1}\boldsymbol{s}(\varrho)[a]^{j+1},\varrho+a\right)\\
&= \Phi\left(\sum_{i=0}^{j}\frac{1}{i!}\partial^i\boldsymbol{s}(\varrho)[a]^i,\varrho+a\right) + \frac{1}{(j+1)!}\partial_s\Phi\left(\sum_{i=0}^{j}\frac{1}{i!}\partial^i\boldsymbol{s}(\varrho)[a]^i,\varrho+a\right)\partial^{j+1}\boldsymbol{s}(\varrho)[a]^{j+1} +\\
&\qquad + O(\|a\|^{2j+2})\\
&= \Phi^{j+1}[a]^{j+1} + O(\|a\|^{j+2}) + \frac{1}{(j+1)!}\partial_s\Phi(\boldsymbol{s}(\varrho),\varrho)\,\partial^{j+1}\boldsymbol{s}(\varrho)[a]^{j+1} + O(\|a\|^{j+2}) +\\
&\qquad + O(\|a\|^{2j+2})\\
&= \Phi^{j+1}[a]^{j+1} + \frac{1}{(j+1)!}\partial_s\Phi(\boldsymbol{s}(\varrho),\varrho)\,\partial^{j+1}\boldsymbol{s}(\varrho)[a]^{j+1} + O(\|a\|^{j+2}).
\end{aligned}$$

Mit Koeffizientenvergleich erhalten wir daraus

$$\Phi^{j+1}[a]^{j+1} = -\frac{1}{(j+1)!}\partial_s\Phi(\boldsymbol{s}(\varrho),\varrho)\,\partial^{j+1}\boldsymbol{s}(\varrho)[a]^{j+1}, \qquad j=0,1,\ldots \tag{4.13}$$

mit

$$\boldsymbol{B}(\boldsymbol{s}(\varrho),\tau) = \partial_s\Phi(\boldsymbol{s}(\varrho),\varrho).$$

Um in (4.12) $\Phi^{j+1}[a]^{j+1}$ für ein gegebenes j berechnen zu können, werden die Terme $\frac{1}{i!}\partial^i\boldsymbol{s}(\varrho)[a]^i$ für $i=0,1,\ldots,j$ benötigt. Für $j=0$ erhält man

$$\Phi(\boldsymbol{s}(\varrho),\varrho+a) = \partial_\varrho\Phi(\boldsymbol{s}(\varrho),\varrho)a + O(\|a\|^2),$$

woraus unter Beachtung von (2.61)

$$\boldsymbol{B}(x,\vartheta,\tau)\partial\boldsymbol{s}(\varrho)[a] = -\Phi^1[a] = -\partial_\varrho\Phi(\boldsymbol{s}(\varrho),\varrho)a = \begin{bmatrix} 0 & -\partial_\tau\boldsymbol{F}(x,\vartheta,\tau) & I_m\\ I_{n+q-m} & 0 & 0\end{bmatrix} a \tag{4.14}$$

folgt. Somit können die Terme $\Phi^{j+1}[a]^{j+1}$ und $\frac{1}{(j+1)!}\partial^{j+1}\boldsymbol{s}(\varrho)[a]^{j+1}$ für $j=1,2,\ldots$ daraus schrittweise unter Beachtung der speziellen Gestalt von Φ und (2.61) gemäß

$$\begin{bmatrix} \boldsymbol{F}\left(\sum_{i=0}^{j}\frac{1}{i!}\partial^i\boldsymbol{s}(\varrho)[a]^i,\tau+a_\tau\right)\\ 0\end{bmatrix} = \Phi^{j+1}[a]^{j+1} + O(\|a\|^{j+2}), \tag{4.15}$$

$$\boldsymbol{B}(x,\vartheta,\tau)\frac{1}{(j+1)!}\partial^{j+1}\boldsymbol{s}(\varrho)[a]^{j+1} = -\Phi^{j+1}[a]^{j+1} \tag{4.16}$$

berechnet werden. Dabei ist $a_\tau = N^T_{n+p+k+l-m,p+k+l-q} M^T_{n+p+k+l,n+p+k+l-m} a$ die τ-Komponente von a. Die Gleichung (4.15) ergibt sich aus (4.12) und die Gleichung (4.16) aus (4.13). Die Taylorkoeffizienten $\Phi^{j+1}[a]^{j+1}$ werden mittels Automatischer Differentiation berechnet, vergleiche die forward-Funktion in [GJU96].

In [GUW97] sind dagegen nur äußere Taylorkoeffizienten gesucht, die auch als äußeren Taylorkoeffizienten einer affin-linearen inneren Funktion interpretiert werden können. Dabei werden als Taylorrichtungen a geeignete Linearkombinationen der verwendeten Ableitungsrichtungen verwendet. Aus den so erhaltenen Taylorkoeffizienten können dann durch Linearkombination alle reinen und gemischten Richtungsableitungen in diese Ableitungsrichtungen bestimmt werden. Der lineare Zusammenhang zwischen Taylor- und Ableitungsrichtungen einerseits und Richtungsableitungen und Taylorkoeffizienten andererseits bleibt bei der hier beschriebenen Vorgehensweise offensichtlich erhalten. Deshalb kann die Berechnung der Richtungsableitungen aus den Taylorkoeffizienten aus [GUW97] unmittelbar verwendet werden.

Für die Berechnung von $\boldsymbol{f}$ und der benötigten Richtungsableitungen von $\boldsymbol{f}$ sind wegen (4.9) und Algorithmus 4.1 gemischte Ableitungen in insgesamt $(n+p+k+l-m+1)$ Richtungen zu berechnen. Die Ableitung höchster Ordnung von $\boldsymbol{\vartheta}$ in $\boldsymbol{f}$ bzw. $f \in \mathbb{R}^{n+p+k+l-m}$ habe die Ordnung $j_{max} \geq 1$. Dann sind Ableitungen bis zur Ordnung $j_{max}+1$ zu berechnen. Wenn die Vorgehensweise aus [GUW97] unmittelbar übertragen wird, sind die Taylorkoeffizienten für $\binom{n+p+k+l-m+1+j_{max}}{j_{max}+1}$ verschiedene Werte von a zu bestimmen. Wenn $j_{max} \notin \{1,2\}$ und $n+p+k+l-m \notin \{1,2\}$, sind mindestens 35 verschiedene Taylorreihen zu berechnen. Diese Anzahl wächst mit wachsendem $n+p+k+l-m$ bzw. j_{max} außerdem stark an. Wenn die charakterisierenden Gleichungen gemäß Unterabschnitt 2.4.2 gewählt werden, enthält f aber nur Ableitungen von $\boldsymbol{\vartheta}$ nach $\boldsymbol{\xi}$ bzw. λ, weshalb die gemischten Ableitungen mit mindestens 2 sonstigen Ableitungsrichtungen nicht gebraucht werden. Dies kann zur Verringerung der Anzahl der benötigten Taylorentwicklungen genutzt werden. Wenn in f nur die Ableitungen von $\boldsymbol{\vartheta}$ nach i Komponenten von $(\boldsymbol{\xi}, \lambda)$ mit $1 \leq i \leq \min\{n+q-m+p, n+p+k+l-m\}$ benötigt werden, kann die Vorgehensweise aus [GUW97] auf alle Kombinationen dieser i Ableitungsrichtungen mit einer weiteren Ableitungsrichtung angewendet werden, was zu $(n+p+k+l-m+1-i) * \binom{i+1+j_{max}}{i}$ Taylorreihen führt. Bei vielen singulären Punkten, z. B. bei solchen mit den Normalformen $\boldsymbol{\vartheta} = \pm\boldsymbol{\xi}^{\ell+1}$, $\boldsymbol{\vartheta} = \pm(\boldsymbol{\xi}^2 \pm \lambda^2)$ bzw. $\boldsymbol{\vartheta} = \pm\boldsymbol{\xi}^{\ell+1} \pm \lambda\boldsymbol{\xi}$, erfolgen dabei die Ableitungen von $\boldsymbol{\vartheta}$ jeweils nur in Richtung einer Komponenten. Deshalb können in diesem Fall die gemischten Ableitungen dieser Komponente mit jeweils einer anderen Komponente gemäß der Vorgehensweise aus [GUW97] berechnet werden. Wenn dabei die gemischten Ableitungen zwischen den i Komponenten bei beiden beteiligten Komponenten berechnet werden, ergeben sich $i*(n+p+k+l-m)*(2+j_{max})$ Taylorreihen. Andernfalls sind $\frac{[2(n+p+k+l-m)-i+1]*i}{2} * (2+j_{max})$ Taylorreihen zu berechnen. Bei der Bestimmung der entsprechenden Richtungsableitung von f ist allerdings die Suche des richtigen Elements in diesem Fall etwas komplizierter. Bei $i = 1$, was z. B. bei Rückkehrpunkten der Fall ist, sind somit wegen den 3 letztgenannten Anzahlen $(n+p+k+l-m) * (2+j_{max})$ Taylorreihen erforderlich.

Beispiel 4.4. Konkret ergeben sich für einige im Abschnitt 2.3 bzw. im Unterabschnitt 2.4.2 behandelten singulären Punkte folgende Werte für die benötigte Anzahl der Taylorreihen. Dabei werden jeweils genau so viele Parameter verwendet, daß diese singulären Punkte stabil gegenüber Störungen von F sind. Das ist gerade dann der Fall, wenn $k+l$ gleich der Kodimension des singulären Punktes ist.

Für die Rückkehrpunkte, d. h. für die singulären Punkte mit der Normalform $\vartheta = \pm\boldsymbol{\xi}^{\ell+1}$ bzw. $g = \pm\xi^{\ell+1} \pm \lambda$ ist $m = n$, $q = p = 1$, $l = 0$, $j_{max} = \ell$ und laut Voraussetzung $\ell = k+1$. Die Anzahl der benötigten Taylorreihen stehen in folgender Tabelle:

$\ell = j_{max} = k+1$	1	2	3	4
$\binom{n+p+k+l-m+1+j_{max}}{j_{max}+1} = \binom{2\ell+1}{\ell+1}$	3	10	35	126
$(n+p+k+l-m)*(2+j_{max}) = \ell(2+\ell)$	3	8	15	24

Für singuläre Punkte mit den Normalformen $\vartheta = \pm(\boldsymbol{\xi}^2 \pm \lambda^2)$, d. h. $\vartheta = \pm(\boldsymbol{\xi}^{\ell+1} \pm \lambda^2)$ mit $\ell = 1$, bzw. $\vartheta = \pm\boldsymbol{\xi}^{\ell+1} \pm \lambda\boldsymbol{\xi}$ mit $\ell = 2, 3, \ldots$ ist $m = n$, $q = p = 1$, $j_{max} = \ell$, $i = 2$ und laut Voraussetzung $\ell = k + l$. Bei den numerischen Berechnungen wurde keine der bisher vorgestellten Varianten von Taylorreihen verwendet. Stattdessen wurden die Ableitungen von ϑ nach $\boldsymbol{\xi}$ und diejenigen nach λ in f getrennt behandelt. Dies kann so interpretiert werden, daß zweimal $i = 1$ gesetzt wurde, wobei einerseits $j_{max} = \ell$ bei den Ableitungen nach $\boldsymbol{\xi}$ und andererseits $j_{max} = 1$ bei den Ableitungen nach λ verwendet wurde. Daraus ergeben sich $(n+p+k+l-m)*(2+\ell)+(n+p+k+l-m)*(2+1) = (1+\ell)*(5+\ell)$ benötigte Taylorreihen. Die Anzahl der benötigten Taylorreihen für die verschiedenen Varianten stehen in folgender Tabelle:

$\ell = j_{max} = k+l$	1	2	3
$\binom{n+p+k+l-m+1+j_{max}}{j_{max}+1} = \binom{2\ell+2}{\ell+1}$	6	20	70
$(n+p+k+l-m+1-i)*\binom{i+1+j_{max}}{i} = \ell*\binom{3+\ell}{2}$	6	20	45
$i*(n+p+k+l-m)*(2+j_{max}) = 2*(\ell+1)*(2+\ell)$	12	24	40
$\frac{[2(n+p+k+l-m)-i+1]*i}{2}*(2+j_{max}) = (2\ell+1)*(2+\ell)$	9	20	35
$(1+\ell)*(5+\ell)$	12	21	32

Für den geflügelten Spitzpunkt mit der Normalform $\vartheta = \pm\boldsymbol{\xi}^3 \pm \lambda^2$ wird für f die erste und zweite Ableitung nach $\boldsymbol{\xi}$, die Ableitung nach λ und die gemischte Ableitung nach $\boldsymbol{\xi}$ und λ benötigt. Es ist wieder $m = n$, $q = p = 1$, $i = 2$, in diesem Fall ist $j_{max} = 2$ und laut Voraussetzung $n+p+k+l-m = 4$. Dann ist $\binom{n+p+k+l-m+1+j_{max}}{j_{max}+1} = \binom{7}{3} = 35$, während $(n+p+k+l-m+1-i)*\binom{i+1+j_{max}}{i} = 3*\binom{5}{2} = 30$ ist. Bei den numerischen Berechnungen wurde eine kompliziertere Variante gewählt, bei der 58 Taylorreihen benötigt werden.

Damit ergibt sich folgender prinzipieller Algorithmus:

Algorithmus 4.5.

1. Berechne die benötigten Werte von h nach den eben beschriebenen Prinzipien.

2. Für jedes derartige h
   berechne $\partial \boldsymbol{s}(\varrho)[a]$ aus (4.14);
   für $j = 1, 2, \ldots, j_{max}$ berechne $\Phi^{j+1}[a]^{j+1}$ mittels Automatischer Differentiation gemäß (4.15) und $\frac{1}{(j+1)!}\partial^{j+1}\boldsymbol{s}(\varrho)[a]^{j+1}$ gemäß (4.16).

3. Berechne aus diesen Werten unter Nutzung der Formeln in [GUW97] die in Schritt 3 des Algorithmus 4.1 benötigten Werte.

Bei diesem zweistufigen Newtonverfahren wird im Gegensatz zum einfachen Newtonverfahren nicht die totale Ableitung von f im aktuellen Punkt benötigt. Alle auftretenden linearen Gleichungssysteme haben außerdem im gleichen Punkt die gleiche Systemmatrix. Deshalb ist offensichtlich der Rechenaufwand geringer als beim einfachen Newtonverfahren. In exakter Arithmetik stimmen jedoch die Iterierten beim einfachen und bei diesem zweistufigen Newtonverfahren überein.

4.3 Numerische Beispiele

Algorithmus 4.1 wurde unter Nutzung des Algorithmus 4.5 im Schritt 3 für die Spezialfälle $m = n$, $p = q = 1$, $l = 0$ in C++ programmiert. Dabei wird vorausgesetzt, daß die Komponenten von f Ableitungen von $\boldsymbol{\vartheta}$ sind. Es wurde ADOL-C verwendet. Dieser Algorithmus wurde auf verschiedene Gleichungssysteme, singuläre Punkte und Startwerte angewendet. Dazu werden lediglich Angaben zur Funktion F, die Startwerte, die Startränderung $\boldsymbol{T} \in \mathbb{R}^{1\times(n+1)}$ und Angaben, welche Ableitungen von $\boldsymbol{\vartheta}$ die Komponenten von f bilden, benötigt. Damit der Algorithmus durchgeführt werden kann, ist die Startränderung dabei so zu wählen, daß die Matrix $\boldsymbol{B}(x, \vartheta, \tau)$ gemäß (4.7) im Startpunkt regulär ist. Weitere Bedingungen braucht $\boldsymbol{T}$ nicht zu erfüllen, da als erster Schritt analog zu Schritt 2 im Algorithmus 4.1 bzw. zum Algorithmus aus Bemerkung 3.28 ein Nullraumvektor von $\partial_{(x,\vartheta)}\boldsymbol{F}(x, \vartheta, \tau)$ im Startpunkt bestimmt wird. Als eigentliche Startränderung $\boldsymbol{T}$ wird dann ein normierter Nullraumvektor verwendet. Die Startwerte sollten so gewählt werden, daß das Newtonverfahren nach einigen Schritten konvergiert. Bei der Funktion F wird eine Funktionsbeschreibung für das Automatische Differenzieren, Angaben zu den Dimensionen und den Parametern benötigt. Für verschiedene singuläre Punkte kann eine unterschiedliche Anzahl von Parametern für das Auftreten dieser singulären Punkte notwendig sein. Um zu vermeiden, daß für jeden singulären Punkt eine extra Funktionsbeschreibung für das Automatische Differenzieren angegeben werden muß, ist jeweils anzugeben, wieviel und welche der Parameter in dieser Funktionsbeschreibung aktuelle Parameter im Sinne von (2.2) sind. Insbesondere ist anzugeben, welche dieser Parameter zu λ bzw. α gehören.

Unter anderem wurden die Beispiele aus [PSS99] nachgerechnet.

Beispiel 4.6 (Beispiel 1 aus [PSS99]). Trigger-Schaltung mit 2 Transistoren, verglei-

che [Pön87], Beispiel 5.2:

$$
\begin{aligned}
F_1(x_1,x_2,x_3,x_4,\lambda,\alpha) &= x_1/R_4 - I_e\big(\chi(\lambda-x_1)+\chi(x_3-x_1)\big) + A_i I_c\big(\chi(\lambda-x_2)+\chi(x_3-x_4)\big)\\
F_2(x_1,x_2,x_3,x_4,\lambda,\alpha) &= A_n I_e\chi(\lambda-x_1) - I_c\chi(\lambda-x_2) + (x_2-x_3)/R_3 - (U_b-x_2)/\alpha\\
F_3(x_1,x_2,x_3,x_4,\lambda,\alpha) &= (1-A_n)I_e\chi(x_3-x_1) - (x_2-x_3)/R_3 + (1-A_i)I_c\chi(x_3-x_4)\\
F_4(x_1,x_2,x_3,x_4,\lambda,\alpha) &= A_n I_e\chi(x_3-x_1) - I_c\chi(x_3-x_4) - (U_b-x_4)/R_5
\end{aligned}
$$

mit

$$
A_i = 0.95,\ A_n = 0.99,\ I_c = 10^{-7},\ I_e = 10^{-9},\ R_3 = 50,\ R_4 = 20,\ R_5 = 5,\ U_b = 6,\\
\lambda \in [0,6],\quad \alpha \in [1,20].
$$

Die Nichtlinearität der Transistoren wird durch

$$
\chi(\psi) = e^{30\psi} - 1
$$

modelliert.

Dabei zeigte sich, daß in Tabelle II zum Beispiel 1 in [PSS99] bei $k = 2$ der Wert für $|\partial^3_{\boldsymbol{\xi}}\boldsymbol{\lambda}|$ falsch gerundet wurde. Der richtige Wert ist $1.60\,e+01$. Wenn in diesem Beispiel α als freier Parameter mit dem Startwert $\alpha^0 = 1.2$ aus [Pön87] betrachtet wird, ergeben sich analog zu den Tabellen 2 und 3 in [Pön87] bzw. I und II in [PSS99] folgende Tabellen.

	Startwert	Zweifacher Rückkehrpunkt
x_1	5.1	5.1551033925
x_2	5.7	5.8081344017
x_3	5.7	5.7645113148
x_4	5.5	5.5691720548
λ	5.7	5.7869539996
α	1.2	1.1294023663

i	norm(F)	$\lvert\partial_{\boldsymbol{\xi}}\boldsymbol{\lambda}\rvert$	$\lvert\partial^2_{\boldsymbol{\xi}}\boldsymbol{\lambda}\rvert$	$\lvert\partial^3_{\boldsymbol{\xi}}\boldsymbol{\lambda}\rvert$
0	2.25 e-01	7.16 e-02	1.49 e-01	7.96 e+00
1	2.89 e-01	2.74 e-01	4.22 e+00	5.47 e+01
2	3.03 e-02	1.60 e-01	7.05 e-01	2.21 e+01
3	2.39 e-02	1.32 e-02	1.89 e-01	9.94 e+00
4	1.35 e-03	1.47 e-03	3.35 e-02	5.44 e+00
5	1.96 e-05	9.08 e-05	2.39 e-03	4.67 e+00
6	1.02 e-07	5.77 e-07	1.61 e-05	4.61 e+00
7	4.54 e-12	2.65 e-11	7.43 e-10	4.61 e+00
8	2.02 e-15	7.33 e-16	4.63 e-15	4.61 e+00
9	7.09 e-15	2.35 e-15	4.88 e-15	4.61 e+00

Dies zeigt die überlineare Konvergenz des Newtonverfahrens.

Beispiel 4.7 (Beispiel 2 aus [PSS99]). Dieses Beispiel entspricht dem Beispiel 5.1 aus

[Pön87]. Dabei wurde das Gleichungssystem

$$\begin{gathered} x_{i-1} - 2x_i + x_{i+1} + h^2(\phi_{i-1} + 10\phi_i + \phi_{i+1})/12 = 0\,, \quad i = 1, 2, \ldots, n, \\ x_0 = x_{n+1} = 0\,, \quad (n+1)h = 2\,, \\ \phi_i := \lambda \exp\left(\frac{x_i}{1+\alpha x_i}\right)\,, \quad i = 0, 1, \ldots, n+1 \end{gathered}$$

für $n = 3, 7, 15, 31$ betrachtet, das aus der Diskretisierung einer Explosionsgleichung entsteht. Die Ergebnisse aus [PSS99], Tabelle III, Werte von $x_{(n+1)/2}$, λ und α im Hysteresepunkt und Anzahl der Iterationen bis zu einer vorgegebenen Genauigkeit, wurden eventuell bis auf die letzte Stelle der Argumentwerte bestätigt.

Beispiel 4.8 (Beispiel 3 aus [PSS99]). Es wird der diskretisierte Brusselator vom Beispiel 3 aus [PSS99] betrachtet. Dabei gilt

$$\begin{aligned} F_{2i-1}(x, \lambda, \alpha) &= x_{2i-3} - 2x_{2i-1} + x_{2i+1} + h^2(\phi_{1,i-1} + 10\phi_{1,i} + \phi_{1,i+1})/12 \\ F_{2i}(x, \lambda, \alpha) &= x_{2i-2} - 2x_{2i} + x_{2i+2} + h^2(\phi_{2,i-1} + 10\phi_{2,i} + \phi_{2,i+1})/12 \end{aligned}, \quad i = 2, \ldots, 41.$$

mit

$$\phi_{j,i} = \phi_j(D_X, D_Y, A_0, L, D_A, D_B, ih, x_{2i-1}, x_{2i}), \quad j = 1, 2,\ i = 1, \ldots, 42,$$

wobei

$$\begin{aligned} \phi_1(D_X, D_Y, A_0, L, D_A, D_B, ih, x_{2i-1}, x_{2i}) = -\frac{L^2}{D_X}((D_B+1)x_{2i-1} - x_{2i-1}^2 x_{2i} - \\ - \Omega(A_0, L, D_A, ih)), \\ \phi_2(D_X, D_Y, A_0, L, D_A, D_B, ih, x_{2i-1}, x_{2i}) = -\frac{L^2}{D_Y}(x_{2i-1}^2 x_{2i} - D_B x_{2i-1}) \end{aligned}$$

und

$$\Omega(A_0, L, D_A, ih) = \frac{A_0}{1+\exp(\omega)}\big(\exp(\omega ih) + \exp(\omega(1-ih))\big), \quad \omega = \frac{L}{\sqrt{D_A}}.$$

gilt. Die Ausdrücke für F_1, F_2, F_{83}, F_{84} ergeben sich analog für $i = 1$ bzw. $i = 42$, wobei x_{-1}, x_0, x_{85} und x_{86} durch A_0, D_B/A_0, A_0, D_B/A_0 ersetzt werden. Diese Werte spiegeln die Neumann-Randbedingungen im stetigen Modell wider. Die Funktion F beschreibt die Diskretisierung eines stetigen Brusselator im Intervall $[0, 1]$ mit der äquidistanten Diskretisierungsschrittweite $h = 1/43$, vergleiche [Gov97a].

Wie in [PSS99] wurde für $\lambda = D_A$, $\alpha = (A_0, L)$ und den festen Werten $D_X = 0.0016$, $D_Y = 0.008$ und $D_B = 5.6$ ein dreifacher Rückkehrpunkt gefunden. Dabei wurden die Werte von x aus [PSS99], Tabelle V, bestätigt.

Beispiel 4.9. Ausgehend von dem dreifachen Rückkehrpunkt aus dem Beispiel 4.8 wurde für $\lambda = L$, $\alpha = (A_0, D_A, D_B)$, $\vartheta = D_A$ bei $D_X = 0.0016$ und $D_Y = 0.008$ ein doppelter Heugabelverzweigungspunkt berechnet. Dieser entspricht P_1^* in [Gov97a]. Analog zum Beispiel 3 in [PSS99] werden in der Tabelle 4.1 wegen der Symmetrie bezüglich des Mittelpunktes von $[0, 1]$ nur die ersten 42 Komponenten von x^* angegeben. Die Para-

Tabelle 4.1:

$x_1 - x_6$	2.340289156	2.427317628	2.413197208	2.394280263	2.481402609	2.362603001
$x_7 - x_{12}$	2.542773378	2.332905381	2.595492251	2.305724872	2.638165218	2.281490138
$x_{13} - x_{18}$	2.669901117	2.260501273	2.690355037	2.242918651	2.699732753	2.228761087
$x_{19} - x_{24}$	2.698757894	2.217912965	2.688607229	2.210139174	2.670821939	2.205106103
$x_{25} - x_{30}$	2.647203981	2.202406625	2.619706625	2.201586963	2.590327288	2.202173527
$x_{31} - x_{36}$	2.561009118	2.203698119	2.533555728	2.205720363	2.509561364	2.207846661
$x_{37} - x_{42}$	2.490356877	2.209745416	2.476970371	2.211158589	2.470100550	2.211909874

Tabelle 4.2:

i	norm(F)	$\partial_{\boldsymbol{\xi}}\boldsymbol{\vartheta}$	$\partial^2_{\boldsymbol{\xi}}\boldsymbol{\vartheta}$	$\partial^3_{\boldsymbol{\xi}}\boldsymbol{\vartheta}$	$\partial^4_{\boldsymbol{\xi}}\boldsymbol{\vartheta}$	$\partial_\lambda\boldsymbol{\vartheta}$	$\partial_{\boldsymbol{\xi}}\partial_\lambda\boldsymbol{\vartheta}$	$\partial^2_\lambda\boldsymbol{\vartheta}$
0	5.19 e-09	2.02 e-10	-1.31 e-09	1.23 e-10	5.41 e-01	-1.52 e+00	-1.19 e+02	-4.13 e+04
1	5.07 e-06	-3.99 e-05	-1.70 e-05	-5.78 e-06	5.01 e-01	-8.56 e-02	-1.11 e+02	-3.80 e+04
2	2.20 e-08	-2.64 e-07	-1.42 e-07	-3.73 e-08	4.99 e-01	-2.42 e-04	-1.10 e+02	-3.78 e+04
3	1.18 e-13	-3.39 e-12	-2.03 e-12	4.16 e-13	4.99 e-01	-2.12 e-09	-1.10 e+02	-3.79 e+04
4	7.54 e-15	9.51 e-14	2.08 e-14	3.11 e-14	4.99 e-01	-2.64 e-11	-1.10 e+02	-3.79 e+04
5	6.94 e-15	-2.81 e-14	4.25 e-15	-1.25 e-14	4.99 e-01	2.77 e-12	-1.10 e+02	-3.79 e+04
6	6.01 e-15	3.04 e-14	6.47 e-14	-1.55 e-14	4.99 e-01	1.75 e-11	-1.10 e+02	-3.79 e+04

meterwerte sind $L^* = 0.202667195161$, $A_0^* = 2.26499861596$, $D_A^* = 0.621345001271$ und $D_B^* = 5.57426279778$. Die Iteration ergibt sich aus Tabelle 4.2.

Für das gleiche Gleichungssystem wurde der geflügelte Spitzpunkt W_1 aus [Gov97a] berechnet. Analog zu Tabelle 4.1 sind die entsprechenden Werte in Tabelle 4.3 dargestellt. Die Parameterwerte sind $L^* = 0.1353905795$, $A_0^* = 4.544434074$, $D_A^* = 0.02097091451$ und $D_B^* = 7.568793411$ bei den festen Werten $D_X = 0.0016$ und $D_Y = 0.008$. Mit $\lambda = L$ und $\vartheta = D_A$ ist außerdem in diesem Punkt $\|F\| \approx 7.76\,e-15$, $\partial_{\boldsymbol{\xi}}\boldsymbol{\vartheta} \approx -1.16\,e-16$, $\partial^2_{\boldsymbol{\xi}}\boldsymbol{\vartheta} \approx 1.66\,e-18$, $\partial^3_{\boldsymbol{\xi}}\boldsymbol{\vartheta} \approx 2.20\,e-04$, $\partial_\lambda\boldsymbol{\vartheta} \approx 2.81\,e-15$, $\partial_{\boldsymbol{\xi}}\partial_\lambda\boldsymbol{\vartheta} \approx -1.94\,e-13$ und $\partial^2_\lambda\boldsymbol{\vartheta} \approx -8.21\,e+00$.

Tabelle 4.3:

$x_1 - x_6$	4.418583844	1.716625108	4.291801019	1.767824077	4.163503744	1.819114009
$x_7 - x_{12}$	4.033486670	1.870423791	3.901906035	1.921605306	3.769253698	1.972438838
$x_{13} - x_{18}$	3.636319314	2.022641264	3.504141406	2.071876828	3.373949628	2.119769992
$x_{19} - x_{24}$	3.247101711	2.165919574	3.125019322	2.209913273	3.009127152	2.251341616
$x_{25} - x_{30}$	2.900799109	2.289810495	2.801314491	2.324951640	2.711825701	2.356430670
$x_{31} - x_{36}$	2.633337722	2.383952630	2.566698301	2.407265211	2.512596809	2.426160032
$x_{37} - x_{42}$	2.471569190	2.440472509	2.444006169	2.450080890	2.430162079	2.454904970

Literaturverzeichnis

[AS97] E. L. Allgower, H. Schwetlick. A general view of minimally extended systems for simple bifurcation points. *Z. Angew. Math. Mech.*, 77:83–97, 1997.

[Att92] B. S. Attili. A direct method for the characterization and computation of bifurcation points with corank 2. *Computing*, 48(2):149–159, 1992.

[Bau57] F. L. Bauer. Das Verfahren der Treppeniteration und verwandte Verfahren zur Lösung algebraischer Eigenwertprobleme. *Z. Angew. Math. Phys.*, 8(3):214–235, 1957.

[Bey84] W.-J. Beyn. Defining equations for singular solutions and numerical applications. In T. Küpper, H. D. Mittelmann, H. Weber (Herausgeber), *Numerical Methods for Bifurcation Problems (Proceedings of the Conference at the University of Dortmund, August 22–26, 1983)*, Seiten 42–56. International Series of Numerical Mathematics 70. Birkhäuser Verlag, 1984.

[Bey90] W.-J. Beyn. Global bifurcations and their numerical computation. In D. Roose, B. De Dier, A. Spence (Herausgeber), *Continuation and Bifurcations: Numerical Techniques and Applications*, Band 313 der *NATO ASI Series C*, Seiten 169–181. Kluwer, Amsterdam, 1990.

[BGMN92] A. Bunse-Gerstner, V. Mehrmann, N. K. Nichols. Regularization of descriptor systems by derivative and proportional state feedback. *SIAM Journal Matrix Analysis and its Applications*, 13(1):46–67, 1992.

[Bjö96] Å. Björck. *Numerical Methods for Least Squares Problems*. SIAM, Philadelphia, 1996.

[Cha84] T. F. Chan. Deflated decomposition of solutions of nearly singular systems. *SIAM J. Numer. Anal.*, 21(4):738–754, 1984.

[DR90] R.-X. Dai, W. C. Rheinboldt. On the computation of manifolds of foldpoints for parameter-dependent problems. *SIAM Journal on Numerical Analysis*, 27(2):437–446, 1990.

[EHM95] L. Elsner, C. He, V. Mehrmann. Minimizing the norm, the norm of the inverse and the condition number of a matrix by completion. *Numerical Linear Algebra with Applications*, 2(2):155–171, 1995.

[FR87] J. P. Fink, W. C. Rheinboldt. A geometric framework for the numerical study of singular points. *SIAM J. Numer. Anal.*, 24(3):618–633, 1987.

[GJU96] A. Griewank, D. Juedes, J. Utke. ADOL-C: A package for the automatic differentiation of algorithms written in C/C++. *ACM TOMS*, 22(2):131–167, 1996. Algor. 755.

[Gov93] W. Govaerts. Computation of Takens-Bogdanov type bifurcations with arbitrary codimension. *SIAM Journal on Numerical Analysis*, 30(4):1121–1133, 1993.

[Gov95] W. Govaerts. Defining functions and nondegeneracy conditions for singularities. Internal Report TWI–95–4, Universiteit Gent, 1995.

[Gov97a] W. Govaerts. Computation of singularities in large nonlinear systems. *SIAM J. Numer. Anal.*, 34(3):867–880, 1997.

[Gov97b] W. Govaerts. Numerical classification of singularities. *Z. Angew. Math. Mech.*, 77 Suppl. 2:S 441–S 444, 1997.

[GR84] A. Griewank, G. W. Reddien. Characterization and computation of generalized turning points. *SIAM J. Numer. Anal.*, 21(1):176–185, 1984.

[GR89] A. Griewank, G. W. Reddien. Computation of cusp singularities for operator equations and their discretizations. *Journal of Computational and Applied Mathematics*, 26(1–2):133–153, 1989.

[GR96] A. Griewank, G. W. Reddien. The approximate solution of defining equations for generalized turning points. *SIAM J. Numer. Anal.*, 33(5):1912–1920, 1996.

[Gri89] A. Griewank. On automatic differentiation. In *Mathematical Programming: Recent Developments and Applications*, Seiten 83–108, Amsterdam, 1989. Kluwer Academic Publishers.

[GS85] M. Golubitsky, D. G. Schaeffer. *Singularities and Groups in Bifurcation Theory*, Band 1. Springer, New York, Berlin, Heidelberg, Tokyo, 1985.

[GUW97] A. Griewank, J. Utke, A. Walther. Evaluating higher derivative tensors by forward propagation of univariate Taylor series. Preprint IOKOMO–09–1997, TU Dresden, 1997. To appear in Computation of Mathematics.

[GVL89] G. H. Golub, C. F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, London, 2nd edition, 1989.

[HVH87] S. Van Huffel, J. Vandewalle, A. Haegemans. An efficient and reliable algorithm for computing the singular subspace of a matrix, associated with its smallest singular values. *J. Comput. Appl. Math.*, 19:313–330, 1987.

[Jan88] V. Janovský. Minimally extended defining conditions for singularities of codim ≤ 2. *Numer. Funct. Anal. Optimiz.*, 9:1309–1349, 1988.

[Jan89] V. Janovský. A note on computing simple bifurcation points. *Computing*, 43:27–36, 1989.

[Jan94] D. Janovská. Numerical treatment of subspace-breaking Takens-Bogdanov points with nonlinear degeneracies. *SIAM Journal on Numerical Analysis*, 31(5):1415–1433, 1994.

[JP95] V. Janovský, P. Plecháč. Numerical analysis of subspace-breaking Takens-Bogdanov points. *IMA Journal of Numerical Analysis*, 15(2):265–290, 1995.

[JS84] A. D. Jepson, A. Spence. Singular points and their computation. In T. Küpper, H. D. Mittelmann, H. Weber (Herausgeber), *Numerical Methods for Bifurcation Problems (Proceedings of the Conference at the University of Dortmund, August 22–26, 1983)*, Seiten 195–209. International Series of Numerical Mathematics 70. Birkhäuser Verlag, 1984.

[JS85] A. D. Jepson, A. Spence. The numerical solution of nonlinear equations having several parameters I: scalar equations. *SIAM Journal on Numerical Analysis*, 22(4):736–759, 1985.

[Kel77] H. B. Keller. Numerical solution of bifurcation and nonlinear eigenvalue problems. In P. H. Rabinowitz (Herausgeber), *Applications of Bifurcation Theory*, Seiten 359–384. Academic Press, New York, 1977.

[KS88] A. Kiełbasiński, H. Schwetlick. *Numerische lineare Algebra. Eine computerorientierte Einführung.* Verlag Harri Deutsch, Thun, Frankfurt/Main, 1988. Auch: Deutscher Verlag der Wissenschaften, Berlin, 1988.

[Kun89] P. Kunkel. Efficient computation of singular points. *IMA J. Numer. Anal.*, 9(3):421–433, 1989.

[Kun91] P. Kunkel. *A unified approach to the numerical treatment of singular points.* Habilitationsschrift, Universität Oldenburg, 1991.

[LST98] R. Lösche, H. Schwetlick, G. Timmermann. A modified block Newton iteration for approximating an invariant subspace of a symmetric matrix. *Linear Algebra Appl.*, 275–276:381–400, 1998.

[Lui97] S. H. Lui. Computation of pseudospectra by continuation. *SIAM J. Sci. Comput.*, 18(2):565–573, 1997.

[Pai74] C. C. Paige. Bidiagonalization of matrices and solution of linear equations. *SIAM J. Numer. Anal.*, 11(1):197–209, 1974.

[Pön85] G. Pönisch. Computing simple bifurcation points using a minimally extended system of nonlinear equations. *Computing*, 35:277–294, 1985.

[Pön87] G. Pönisch. Computing hysteresis points of nonlinear equations depending on two parameters. *Computing*, 39(1):1–17, 1987.

[Pön90] G. Pönisch. A two-stage NEWTON-like method for computing simple bifurcation points of nonlinear equations depending on two parameters. In *Numerical analysis and mathematical modelling*, Seiten 121–133. Banach Center Publications 24, 1990.

[Pön91] G. Pönisch. An indirect method for computing origins for Hopf bifurcation in two-parameter problems. *Computing*, 46(4):307–320, 1991.

[Pön92] G. Pönisch. An indirect approach to computing origins of Hopf bifurcation and its application to problems with symmetry. In E. Allgower, K. Böhmer, M. Golubitsky (Herausgeber), *Bifurcation and symmetry*, Seiten 285–294. International Series of Numerical Mathematics 104. Birkhäuser Verlag, 1992.

[Pön94] G. Pönisch. Two-stage methods for computing singular points of unterdetermined nonlinear equations. *Z. Angew. Math. Mech.*, 74(6):T664–T665, 1994.

[PS81] G. Pönisch, H. Schwetlick. Computing turning points of curves implicitly defined by nonlinear equations depending on a parameter. *Computing*, 26(2):107–121, 1981.

[PSS97a] G. Pönisch, U. Schnabel, H. Schwetlick. Computation of multiple pitchfork bifurcation points. *Z. Angew. Math. Mech.*, 77 Suppl. 2:S 449–S 452, 1997.

[PSS97b] G. Pönisch, U. Schnabel, H. Schwetlick. Computing multiple pitchfork bifurcation points. *Computing*, 59:209–222, 1997.

[PSS98] G. Pönisch, U. Schnabel, H. Schwetlick. Computing multiple turning points. *Z. Angew. Math. Mech.*, 78 Suppl. 3:S 1039–S 1040, 1998.

[PSS99] G. Pönisch, U. Schnabel, H. Schwetlick. Computing multiple turning points by using simple extended systems and computational differentiation. *Optimization Methods and Software*, 10(4):639–668, 1999.

[Rhe93] W. C. Rheinboldt. On the sensitivity of solutions of parametrized equations. *SIAM Journal on Numerical Analysis*, 30(2):305–320, 1993.

[RR86] P. J. Rabier, G. W. Reddien. Characterization and computation of singular points with maximum rank deficiency. *SIAM Journal on Numerical Analysis*, 23(5):1040–1051, 1986.

[Rut69] H. Rutishauser. Computational aspects of F. L. Bauer's simultaneous iteration method. *Numerische Mathematik*, 13(1):4–13, 1969.

[Rut70] H. Rutishauser. Handbook series linear algebra. Simultaneous iteration method for symmetric matrices. *Numerische Mathematik*, 16(3):205–223, 1970.

[Sch79] H. Schwetlick. *Numerische Lösung nichtlinearer Gleichungen.* Deutscher Verlag der Wissenschaften, Berlin, 1979. Auch: R. Oldenbourg Verlag, München-Wien, 1979.

[Sch94] U. Schnabel. Minimale Systeme zur Charakterisierung singulärer Punkte. Diplomarbeit, Technische Universität Dresden, 1994.

[Sch98] U. Schnabel. Über Ljapunov-Schmidt-reduzierte Funktionen und numerisch auswertbare Analoga. Preprint IOKOMO–10–1998, TU Dresden, 1998.

[Sey79] R. Seydel. Numerical computation of branch points in nonlinear equations. *Numerische Mathematik*, 33:339–352, 1979.

[She97] Y.-Q. Shen. Computation of a simple bifurcation point using one singular value decomposition nearby. *Computing*, 58(4):335–350, 1997.

[SL97] H. Schwetlick, R. Lösche. A generalized inverse iteration for computing simple eigenvalues of nonsymmetric matrices. Preprint IOKOMO-07-97, TU Dresden, 1997. To appear in *Z. Angew. Math. Mech.*

[SPJ99] U. Schnabel, G. Pönisch, V. Janovský. Reduced functions characterizing singular points and their relations. Preprint IOKOMO–04–1999, TU Dresden, 1999.

[SRS69] H. R. Schwarz, H. Rutishauser, E. Stiefel. *Numerik symmetrischer Matrizen.* BSB B. G. Teubner Verlagsgesellschaft, Leipzig, 1969. Auch: Verlag B. G. Teubner, Stuttgart 1968.

[SS97a] S. Schleiff, H. Schwetlick. Characterization and computation of period doubling points by minimally extended systems. *Optimization Methods and Software*, 8:1–24, 1997.

[SS97b] H. Schwetlick, U. Schnabel. Iterative computation of the smallest singular value and the corresponding singular vectors of a matrix. Preprint IOKOMO–06–97, TU Dresden, 1997.

[Ste69] G. W. Stewart. Accelerating the orthogonal iteration for the eigenvectors of a Hermitian matrix. *Numerische Mathematik*, 13(4):362–376, 1969.

[Ste81] G. W. Stewart. On the implicit deflation of nearly singular systems of linear equations. *SIAM J. Sci. Statist. Comp.*, 2:136–140, 1981.

[SW99] U. Schnabel, A. Walther. Berechnung singulärer Punkte: Newtonverfahren und Automatische Differentiation. Preprint IOKOMO–05–1999, TU Dresden, 1999.

[Wil65] J. H. Wilkinson. *The algebraic eigenvalue problem.* Clarendon Press, Oxford, 1965.

Erklärung gemäß §5, Absatz (1), Punkt 5 a), b) der Promotionsordnung

Versicherung

Hiermit versichere ich, daß ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht. Die Arbeit wurde bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt.

Die vorgelegte Dissertation wurde am Institut für Numerische Mathematik im Fachbereich Mathematik an der Technischen Universität Dresden unter wissenschaftlicher Betreuung von Herrn Prof. Dr. rer. nat. habil. H. Schwetlick angefertigt.

Dresden,

HSSS AdminTools (c) 2001, last visited: Thu Aug 30 14:59:37 GMT+02:00 2001